US-00059-TS(1)

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２０００年 ６月３０日

出願番号
Application Number: 特願２０００－１９９７３８

出願人
Applicant (s): 松下電器産業株式会社

CERTIFIED COPY OF PRIORITY DOCUMENT

２０００年１２月２２日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号 出証特２０００－３１０８２５０

【書類名】　特許願
【整理番号】　2030724007
【提出日】　平成12年 6月30日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/27
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地
　　松下電器産業株式会社内
　【氏名】　飯塚　泰樹
【特許出願人】
　【識別番号】　000005821
　【氏名又は名称】　松下電器産業株式会社
【代理人】
　【識別番号】　100082692
　【弁理士】
　【氏名又は名称】　蔵合　正博
　【電話番号】　03-5210-2681
【選任した代理人】
　【識別番号】　100081514
　【弁理士】
　【氏名又は名称】　酒井　一
　【電話番号】　03-5210-2681
【手数料の表示】
　【予納台帳番号】　013549
　【納付金額】　21,000円
【先の出願に基づく優先権主張】
　【出願番号】　平成11年特許願第373272号
　【出願日】　平成11年12月28日

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9004843

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　単語分割方式及び装置

【特許請求の範囲】

【請求項１】　文を単語に分割する単語分割方式において、単語分割されていない文書データから文字結合度を文字間接続確率という形で統計的に計算し、計算した文字間接続確率を分割対象の文に適用し、文字間接続確率が低い部分で分割することで文を単語に分割することを特徴とする単語分割方式。

【請求項２】　文を単語に分割する単語分割装置において、文の集りである文書データを蓄える文書データ蓄積手段と、単語に分割されていない文書データから文字間接続確率を計算する文字間接続確率計算手段と、計算した文字間接続確率の値を蓄える確率テーブル記憶手段と、計算した文字間接続確率を用いて文を単語単位に分割する文字列分割手段と、文書入力手段と文書出力手段を備えることを特徴とする単語分割装置。

【請求項３】　文を単語に分割する単語分割方式において、単語分割されていない文書データから文字結合度を文字間接続確率という形で統計的に計算する際、この文字結合度として複数文字からなるある文字列が出現した後にある文字が出現する確率として計算し、この確率を分割対象の文に適用し、文字間接続確率が低い部分で分割することで文を単語に分割することを特徴とする単語分割方式。

【請求項４】　文を単語に分割する単語分割方式において、単語分割されていない文書データから文字結合度を文字間接続確率という形で統計的に計算する際、この文字結合度として複数文字からなるある文字列が出現した後に複数文字からなるある文字列が出現する確率として計算し、この確率を分割対象の文に適用し、文字間接続確率が低い部分で分割することで文を単語に分割することを特徴とする単語分割方式。

【請求項５】　文を単語に分割する単語分割方式において、文書データから文字結合度としての文字間接続確率を統計的計算する際、複数文字からなるある文字列が出現した後にある文字が出現する確率と複数文字からなるある文字列が出現する前にある文字が出現する確率とから、複数文字からなるある文字列が出現

した後に複数文字からなるある文字列が出現する確率を計算することを特徴とした前記請求項４の単語分割方式。

【請求項６】　文を単語に分割する単語分割方式において、単語分割されていない学習用文書データから文字結合度としての文字間接続確率を統計的に計算し、計算した文字間接続確率を分割対象の文に適用し、もし分割対象の文の中に文字間接続確率として計算した以外の文字組み合わせが出現した場合は既計算文字間接続確率から目的とする確率を推定し、文字間接続確率が低い部分で分割することで文を単語に分割することを特徴とする単語分割方式。

【請求項７】　文字間接続確率による単語分割において、分割するかどうかの判断の基準となる確率値の閾値を、分割後の平均単語長をもとに動的に決定することを特徴とする請求項１記載の単語分割方式。

【請求項８】　日本語の文の単語分割において、文字間接続確率の他に、漢字や平仮名といった文字の種類が変化する点で単語が分割されやすいという特徴を併用することで単語分割点を決定することを特徴とした前記請求項１の単語分割方式。

【請求項９】　文字間接続確率の他に、カッコなどの記号部分では必ず単語が分割されるということを併用することで単語分割点を決定することを特徴とした前記請求項１の単語分割方式。

【請求項１０】　文を単語に分割する単語分割方式において、単語分割されていない文書データから文字結合度を文字間接続確率という形で統計的に計算して記憶した後、単語辞書を用いて文字列を単語単位に分割する時に単語分割の解の候補が複数あるときは先に計算した文字間接続確率に注目し、分割の解の候補のうち単語分割点における文字間接続確率が小さいものを選択することによって、文を単語に分割することを特徴とする単語分割方式。

【請求項１１】　請求項１０に記載の単語分割方式において、単語分割の解の候補が複数ある時は、夫々の候補の単語分割点における文字間接続確率の和をその解の候補のスコアとし、分割の解同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語分割方式。

【請求項１２】　請求項１１に記載の単語分割方式において、単語分割の解の

候補が複数ある時は、夫々の候補の単語分割点における文字間接続確率の積をその解の候補のスコアとし、分割の解同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語分割方式。

【請求項１３】　請求項１０に記載の単語分割方式において、全ての文字位置についてその文字位置の前が単語分割点だったら文字間接続確率を、単語分割点でなければ定数を与え、それらの確率と定数の和をもって単語分割の解の候補のスコアとし、分割の解の候補同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語分割方式。

【請求項１４】　請求項１０に記載の単語分割方式において、全ての文字位置についてその文字位置の前が単語分割点だったら文字間接続確率を、単語分割点でなければ定数を与え、それらの確率または定数の積をもって単語分割の解の候補のスコアとし、分割の解同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語分割方式。

【請求項１５】　文を単語に分割する単語分割方式において、単語分割されていない文書データから文字結合度を文字間接続確率という形で統計的に計算して記憶した後、単語辞書を用いて文字列を単語単位に分割する時に、辞書に無い文字列を未知語として分割候補に含めることで分割候補を作成し、単語分割の解の候補が複数あるときは先に計算した文字間接続確率に注目し、分割の解の候補のうち単語分割点における文字間接続確率が小さいものを選択することによって、文を単語に分割することを特徴とする単語分割方式。

【請求項１６】　請求項１５に記載の単語分割方式において、文字列のある文字位置で終了する単語が辞書にあるがその次の文字位置から開始される単語が辞書に無い時、そこから始まるｎ文字以上ｍ文字以下の文字列を全て未知語として扱うことで未知語の候補を網羅することを特徴とする単語分割方式。

【請求項１７】　請求項１５に記載の単語分割方式において、辞書にある単語には定数スコアを与え、未知語推定された単語にはそれより高い定数のスコアを与え、単語分割の解の候補ごとに単語のスコアの和と分割点における文字間接続確率の和を取ってこれを解の候補のスコアとし、分割の解同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語

分割方式。

【請求項１８】　請求項１５に記載の単語分割方式において、辞書にある単語には定数スコアを与え、未知語推定された単語にはそれより高い定数のスコアを与え、単語分割の解の候補ごとに単語のスコア積と分割点における文字間接続確率の積を取ってこれを解の候補のスコアとし、分割の解同士をこれらのスコアで比較し、最も低いスコアを取るものを解として選択することを特徴とする単語分割方式。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、電子計算機を利用した機械翻訳や大量文書検索、テキスト自動要約等を実施する自然言語処理システムの前処理・解析部における方式と装置に関し、特に、文を効率的に単語単位に分割できるようにしたものである。

【０００２】

【従来の技術】

以後の本発明の説明において、単語は文字の列（文字列）から構成されたもので、文字が組み合わさって意味を形成する単位とする。文（あるいは文章）は単語の列から構成されているものであり、結果として文字列で表される。文書とは文が複数集まってまとまりを作った単位であるとする。

【０００３】

日本語や中国語など単語を分けて書かない言語を膠着語という。膠着語では、言語の知識を持たない者がその字面だけ見ると、文は長い文字列であって、単語の境界をみつけることができない。

【０００４】

機械翻訳や自動要約といった自然言語処理システムにおいては、その最初の段階として文の解析が必要になる。日本語のような膠着語では、単語への分割が最初の解析に相当する。

【０００５】

また文書検索システムでは、例えば「今月の東京都議会」という文字列の「東

京都議会」という語を単語の概念を使わずに文字列検索できるようにしてしまうと、「東京」で検索した場合でも「京都」で検索した場合でもヒットしてしまうことになる。検索語が「京都」だった時に「東京都」がヒットしてしまうのは、本来望まない結果であることから検索ノイズと呼ばれる。こうした検索ノイズを減らすためには、検索対象文書の文を単語に分割しておく必要がある。

【０００６】

このような単語分割処理には、通常は辞書を用いた形態素解析処理が使われる。形態素解析では、解析用の辞書を用いて文を単語へ分割するが、形態素解析の精度はこの辞書がどれだけ整っているかに依存する。

【０００７】

一方、近年、文書中の文字列や文字の出現といったものを統計的に調べて、処理に必要な情報を得るという提案がなされている。これは例えば、既に単語に分割されている文書から、ある単語（または単語列）の次にどのような単語が出現しやすいかというものを確率として計算し、形態素解析の時にこの情報を使って解の候補を絞るというものである。（参考文献：「単語と辞書」松本祐治　他著、岩波書店　１９９７年）確率の計算には、単語Ｎグラムという単語Ｎ個組が使われる。ＮグラムはＮ－１個の単語の次に、ある単語が出現する確率を計算するもので、この確率計算はマルコフモデルとも呼ばれており、音声認識の単語推定などにも応用されている。ただし単語Ｎグラムは単語の接続可能性を計算するものであって、辞書にない単語を類推するものではない。また、この単語Ｎグラムの学習には既に単語に分割されている大量の文書が必要であるが、このような文書は機械的に作ることはできず、人により検査しながら作成する必要があるため、用意するには大きなコストがかかる。

【０００８】

この統計処理の考え方を用いるものとして、単語のＮグラムではなく、文字に着目した文字Ｎグラムがある。文字Ｎグラムは、（Ｎ－１）個の文字列の後にどのような文字が続くかの確率を計算したものである。

【０００９】

この文字Ｎグラムを応用し、文書中に出現する単語になりえそうな文字列の出

現頻度を網羅的に調べて、その文字列前後の文字接続がどれほど散らばっているかを分散という尺度で計算することで単語や慣用句を収集する方法が特開平９－１３８８０１に開示されている。

【００１０】

論文「正規化頻度による形態素境界の推定」（情報処理学会　自然言語処理研究会 NL-113-3 1996）では、Ｎグラムの出現頻度を正規化計算することで、辞書を用いずに文を単語単位へ分割する手法が提案されている。

【００１１】

辞書を用いない形態素解析は特開平１０－３２６２７５、特開平１０－２５４８７４などにも開示されている。特開平１０－３２６２７５は、文字Ｎグラムを使って、文字列の部分連鎖確率と単語分割点との関係をテーブルに記憶しておき、そのテーブルを使って単語分割を行うものである。テーブルの作成には、あらかじめ単語単位に分割された文（または文書）を用意し、その文書から計算機により自動学習を行わせる。特開平１０－２５４８７４も同様に単語単位に分割された文書からあらかじめ学習をする必要がある。

【００１２】

【発明が解決しようとする課題】

しかしながら、言語には常に新しい単語が生まれるものであるため、形態素解析用辞書は常にメンテナンスが必要である。また、対象とする文書によって単語の使われ方が違うこともあり、対象とする文書を変更する度に辞書を調整しなければいけない。そして、どれだけ注意していても形態素解析において未知語、すなわち辞書に載っていない単語に遭遇する可能性は否定できず、未知語の出現により形態素解析の精度が低下することがある。

【００１３】

辞書を使わないかわりに統計的処理を用いるものとして、前記の通り特開平１０－３２６２７５などがある。しかしこれらは事前に、単語単位に分割された文書を読ませることでシステムを訓練（自動学習）しておく必要がある。単語単位に分割された文書を用意するためには、人手で文を分割しておくか、または既存の形態素解析システムの出力結果を用いる。だが人手で文を分割するのは多大な

コストが必要であり、文書分野や時代ごとに大量の文を分割して用意することは難しい。そして時代の変化とともに変わっていく言語について、常に大量の分割済み文書を作成し続けなければならず、辞書の整備以上に大変な作業となる。また既存の形態素解析の出力結果を用いた場合、既存の形態素解析における解析失敗部分をそのまま学習してしまい、既存の形態素解析を越える精度は期待できなくなる。

【００１４】

本発明は上記の従来技術の課題を解決するためのものであり、基本的に辞書や単語単位に分割された大量の訓練用の文を必要とせず、文を単語へ分割することができる単語分割方式を提供し、また、その方式を実施する装置を提供することを目的としている。

【００１５】

【課題を解決するための手段】

この目的を達成するために本発明は、単語に分割されていない文書から文字間の結合度を文字間接続確率という形で計算し、この文字間接続確率を使うことで文を単語単位へ分割するものである。これにより、辞書を使わず、また単語分割後文書を学習する必要もなく、文を単語に分割するという効果を奏するものである。

【００１６】

【発明の実施の形態】

（実施の形態１）

以下、本発明の実施の形態について説明する。まず前提となる言語の性質を説明する。文字の出現確率に注目してみる。一般に単語を構成する文字列は、全ての文字の組み合わせの単語が存在するわけではないので、文字の出現は等確率ではない。すなわちある言語の文字の種類をＫ種類とすると、もし単語を構成する文字が等確率で使われているなら、Ｍ文字からなる単語の種類はＫのＭ乗個存在することになる。しかし、実際には語彙数はそれほど多くない。

【００１７】

以後の説明では、膠着語として日本語を例にして説明する。日本語の日常生活

で通常使われる文字の種類は、約６千である。この数は、今日の一般的なコンピュータで扱える（ＪＩＳで規定された）文字の種類数から類推したものである。ここで、日本語２文字の単語について考える。もし全ての文字の組み合わせの単語が存在するなら、６０００の２乗＝３千６百万の単語が存在することになる。日本語にはこの他に３文字や４文字の単語も存在するからさらに多くの単語が存在することになる。しかし日本語の総語彙数はたかだか数十万と考えられる。この数は、岩波書店の広辞苑等、日本語辞書の語彙数が２０万から３０万の間にあることから推測したものである。　これについては、「自然言語処理」（長尾真編　岩波書店　１９９６年出版）の第２章第１節「言語の統計」にも述べられている通りで、一般に文字の出現には偏りがあるものとされている。

【００１８】

次に本発明の原理につて説明する。ある文字「ａ」の後に別の文字「ｂ」が続く確率は、もし前記の偏りがなければ、すなわちどんな文字も等確率で出現して単語を形成するなら、言語を構成する文字の種類Ｋの逆数（日本語の場合約６千分の１）になる。しかし実際には偏りがあるのでそうはならない。具体例で説明する。ある文字列「衆議院」が単語であったとしよう。すると、文字単位での接続確率を「衆」の後に「議」が続く条件付き確率Ｐ(･議｜衆)･とした時、この確率は日本語全体を調べてみるなら６千分の１より大きくなるはずである。同様に文字列「衆議」の後に文字「院」が続く条件付き確率Ｐ(･院｜衆議)･の場合は前２文字が与えられることから、さらに高い確率を示すはずである。一方で、存在しない単語（文字の組み合わせ）と思われる「衆ぴ」などが出現する確率Ｐ(･ぴ｜衆)･は、限りなく０に近くなるはずである。

【００１９】

一方、文を構成する単語は、かなり自由な組み合わせが可能である。例えば「これは数学の本だ」「これは音楽の本だ」は両方とも文であるが、「数学」「音楽」の部分は自由な単語が接続できる。つまり単語を構成する文字列「これは」の後に別の単語の文字「数」が続く条件付き確率Ｐ(･数｜これは)･は、単語を構成する文字間の接続確率よりも低くなるはずである。この文字間接続確率は、文字の間の結合度と解釈できる。そしてこれが計算できれば、それを基に文字列（

文）を単語単位へ分割することができる。

【００２０】

文字間の接続確率は、文をある程度の量、つまりある程度の量の文書を集めることができれば、そこから統計的に調べて計算することができる。すなわち、文書データベースを構築するような状況ならば、データベースに登録する文書から文字間接続確率を統計的に調べて計算することができる。この計算値は日本語全体について調べた場合の確率値とは違うものであろうが、近似できるものであり、しかもその確率値を調べた文書、あるいは類似の文書の分割に適用するのに適した性質を持つ。

【００２１】

本発明では以上の原理を用いることで、単語分割されていない文書から文字間接続確率を調べ、その文字間接続確率を使うことで、文書を辞書を用いることなく単語単位へと分割する。以下、本発明の実施の形態について図面を用いて説明する。

【００２２】

（実施の形態１）

図１は本発明の実施の形態１における単語分割処理方式を説明するフロー図である。図２は本実施の形態１における文字列分割装置の構成を示すブロック図であり、処理対象文書を電子化された状態で入力するための文書入力手段２０１と、入力した文書を蓄えておく文書蓄積手段２０２と、文書から文字間接続確率を計算するための文字間接続確率計算手段２０３と、計算した確率を記録しておくための確率テーブル格納手段２０４と、計算した文字間接続確率を使って文書を単語単位に分割するための単語分割手段２０５と、処理結果の文書を出力する出力手段２０６を備えている。

【００２３】

以上のように構成された文字列分割装置について、その処理動作を図１を用いて説明する。

ステップ１０１：文書入力手段２０１から入力されたデータは、まず文書蓄積手段２０２に蓄えられる。

ステップ１０２：このデータから、文字間の接続確率を２０３の文字間接続確率計算手段が計算し、計算結果を確率テーブル格納手段２０４に蓄える。計算方法の詳細は後述する。

ステップ１０３：文書蓄積手段２０２に蓄えられたデータについて、確率テーブル格納手段２０４に蓄えられた確率値を用いることで、文字間の接続確率を調べ、その確率が低い所で分割をし、

ステップ１０４：分割された文を出力手段２０４から出力する。

【００２４】

以上のように本実施の形態１における文字列分割装置は、処理対象文書から文字間接続確率を計算し、計算した確率を使って対象文書を単語単位へ分割することができる。

【００２５】

以下、図１のステップ１０２の詳細について説明する。　本発明の第一の実施の形態においては、文字Ｃi-1･と文字Ｃi･の間の文字間接続確率は、文字列Ｃ1.．･.Ｃi-1･の後に次の文字Ｃi･が続く条件付き確率で表現することにする。これを式にすると次のように書ける。

【００２６】

【数１】

$$P(C_i|C_1C_2...C_{i-1}) \qquad (1)$$

【００２７】

しかしこれは計算量が大きく記憶空間が大量に必要になる。（１）式のような単語列や文字列の条件付き確率は、一般にはＮグラム（Ｎ＝１，２，３，４，・・，）、と呼ばれる文字Ｎ個組で近似する。文字Ｎグラムよる条件付き確率とは、そのＮ－１個の文字列Ｃi-N+1...Ｃi-1･という文字列が続いたという条件のもとで文字Ｃi･が出現する確率である。すなわち、Ｎグラムの１番目からＮ－１番目の文字列が続いたという条件のもとでＮ番目の文字が出現する確率である。こ

れは（２）式のように書くことができる。

【００２８】

【数２】

$$P(C_i|C_{i-N+1}\cdots C_{i-1}) \qquad (2)$$

【００２９】

Ｎグラムの確率は、文字列Ｃ1･Ｃ2･．．Ｃm･が調べようとするデータ中に出現する回数をCount･（Ｃ1･Ｃ2･．．Ｃm･）とすると、

【００３０】

【数３】

$$P(C_i|C_{i-N+1}\cdots C_{i-1}) = \frac{Count(C_{i-N+1}\cdots C_i)}{Count(C_{i-N+1}\cdots C_{i-1})} \qquad (3)$$

と推定できる（参考文献：「単語と辞書」（松本祐治　他　著　岩波書店　１９９７年）。

【００３１】

なおＮグラム計算の時には、計算する文字列（文）の前後にＮ－１文字の特殊な記号を付与して計算するのが一般的である。（参考文献：同書）これは、一般のＮグラムは文の先頭の文字の確率や、文の最後の文字の確率を、特殊記号を含めたＮグラムにより計算するからである。Ｎ＝３の例で説明するなら、「これは本だ」という文字列の３グラムを作成するためには、特殊記号をここでは＃で表現することとして、「＃＃これは本だ＃＃」のような文字列を作成してからＮグラムを作成する。すると「＃＃こ」「＃これ」「これは」「れは本」「は本だ」「本だ＃」「だ＃＃」の７つの３グラムを作ることになる。

【００３２】

一方、本発明の第１の実施の形態においては一般的なＮグラムの計算とは違い

、計算する文字列の前後にＮ－１文字の特殊な記号を付与せず、計算する文字列の前にだけＮ－２文字（ただしＮ－２≧０とし、Ｎ＝１の時は０とする）の特殊記号を付与する。これは文の後については、文の最後の文字の後は必ず単語として切れるのであるから、文の最後の文字とその後の特殊記号との接続確率を計算する必要がないからである。また文の前については、文の先頭が単語区切であることは自明であるのでＮ－１個の特殊記号は必要なく、文頭１文字と次の１文字との接続確率を計算するために特殊記号Ｎ－２個を含むＮグラムを作成する必要がある。

【００３３】

先の「これは本だ」の３グラムの例で言うなら、文の先頭が単語区切であるのは自明なので、「＃＃こ」によって「＃＃」と「こ」の接続確率を計算する必要は無いが、「＃これ」によって「＃こ」と「れ」の接続確率を計算する必要があるので、文の前に必要な特殊記号の数はＮ－２個となる。同様に「本だ＃」によって「本だ」と「＃」の接続確率を計算する必要はないので、文末に必要な特殊記号の数は０個となる。

【００３４】

式（３）を計算し、文字Ｎ個組とともに計算結果を図２の確率テーブル格納手段２０４に記録することが図１のステップ１０２に相当する。確率テーブル格納手段２０４は、例えば図４（ｄ）のように、Ｎ文字組とその確率値が格納されるものであるが、文字組で検索しやすく記憶容量も小さくするために、適切な構造を用いて実現されているものとし、ここではその構造を限定しない。

【００３５】

ステップ１０２の計算手順は、例えば図３に示す手順で実現できる。

ステップ３０１：文書を構成する文ごとに、文の前に文頭を表現する特殊記号をＮ－２個付与する。

ステップ３０２：Ｎ－１グラム統計を作成する。すなわち、対象文書の中に出現した全ての文字Ｎ－１個組について、それが何回出現しているかを調べた表を作成する。一般にＮグラムの統計を調べる方法は、（参考文献：「言語情報処理」長尾真　他著、岩波書店　１９９８年）などに述べられているが、単純には文

字の種類ＫのＮ乗を表現できるテーブルを用意し、そこに出現数をカウントして行くか、あるはい文書から全てのＮ文字組みを取り出しそれをソートして同じものの出現回数をカウントすれば計算できる。

ステップ３０３：Ｎグラム統計を作成する。すなわち、対象文書の中に出現した全ての文字Ｎ個組について、それが何回出現しているかを調べた表を作成する。これはステップ３０２と同様である。

ステップ３０４：Ｎグラム統計の夫々の文字Ｎ個組文字列について、その出現回数をＸとする。同文字Ｎ個組文字列について、その１番目からＮ－１番目の文字列の出現回数をステップ３０１で作成したＮ－１グラム統計から調べ、これをＹとする。Ｘ／Ｙにより式（３）の値を計算し、この値を確率テーブル格納手段２０４に記録する。

【００３６】

以上により式（３）の値が計算できるが、Ｎ－１グラムを作成しない方法も存在する。ＮグラムはＮ－１グラム文字列を含むことから、Ｎグラムを作っておけばＮ－１グラムの出現頻度も簡単に計算できるからである。

【００３７】

以下、文字間接続確率の具体的な計算例を示す。全文書として文字列「ａｂａａｂａ」だけが与えられた場合を例とし、ここから文字間接続確率をＮ＝３のＮグラム（３グラム）で計算する。まずステップ３０１として、文（文字列）の前に文頭文末を表現する特殊記号をＮ－２（３－２＝１）個付与する。この様子を図４（ａ）に示す。特殊記号としてここでは＃を付けているが、実際には文書に現れない記号を付けるものとする。次にステップ３０２として、２グラムの統計、すなわち文字２個組の出現回数を調べる。その結果が図４（ｂ）のようになる。同様にステップ３０３として３グラムの、文字３個組の出現回数を調べ、図４（ｃ）を得る。ステップ３０４として、図４（ｂ）と図４（ｃ）から、文字３個組についての式（３）の値を計算し、図４（ｄ）を得る。以上が図１のステップ１０２の詳細説明である。

【００３８】

以下では図１のステップ１０３の詳細を説明する。ステップ１０３はステップ

１０２で計算した文字間接続確率の表を使って、処理対象の文を構成する文字のそれぞれの部分の接続確率を調べ、分割をする過程である。その計算手順を図５に示す。本発明の第一の実施の形態においては、δを閾値とし、この値はあらかじめ決められているものとする。

ステップ５０１：文書から文を一つ選択する。

ステップ５０２：図３のステップ３０１と同様に、文の前に文頭を表現する特殊記号をＮ－１個付与する。

ステップ５０３：ポインタを文の前に付けた特殊記号の一文字目に移動する。

ステップ５０４：ポインタ位置から始まるＮ文字について、ステップ１０２で計算した文字間接続確率を調べる。

ステップ５０５：もしその確率が、あらかじめ決められた閾値δ未満だったら、ポインタ位置を１文字目とした時のＮ－１文字目とＮ文字目の間は単語分割点だったものと推定され、よってそこで分割を行う。もしその確率が閾値δ以上だったら、そこは単語分割点ではないので分割を行わない。

ステップ５０６：ポインタを一文字進める。

ステップ５０７：ポインタを１文字目とした時のＮ文字目が文末文字を越える場合、文は終了したものとして、ステップ５０８へ。そうでなければステップ５０４へジャンプする。

ステップ５０８：文書から次の文を選択する。

ステップ５０９：次の文が無ければ終了。そうでなければステップ５０２へ進む。

【００３９】

以上により分割点を発見する。以下、具体的計算例を、先に示した図４の文字列「ａｂａａｂａ」のＮ＝３の場合で示す。既に図４（ｄ）の文字３個組の接続確率は計算されているものとする。閾値としてδ＝０．７が与えられているものとする。まずこの例の場合では文が一つしかないので、ステップ５０１で「ａｂａａｂａ」が選択され、ステップ５０２で文の前に特殊文字が付けられることで図４（ａ）と同じ状態になる。次にステップ５０３でポインタを移動させた状態が図４（ｅ）である。ここから３文字、すなわち「＃ａｂ」の確率を図４（ｄ）

のテーブルで探すと１．０であり、これは閾値δ＝０．７より大きいので、「＃ａ」と「ｂ」の間は分割されない。以下同様にステップ５０４からステップ５０６を繰り返すことで、文字列のそれぞれの点での接続確率が調べられ、この値をもって単語分割点を決定することができる。この様子を図４（ｆ）に示す。この例では、単語分割された結果は「ａｂａ／ａｂａ」となる。

【００４０】

もう一つ別の例として、同様に日本語の単純な文字列「にわにわにわ」（庭には二羽）を計算したのが図６である。文頭特殊記号を付与したものが図６（ａ）、２グラムと３グラムの出現回数はそれぞれ図６（ｂ）（ｃ）に計算され、そこから３グラムの文字間接続確率は図６（ｄ）のようになる。これを元の文字列にあてはめていくと、図６（ｅ）のようになり、閾値δ＝０．７とすることで、結果として「にわ／にわ／にわ」と分割される。

【００４１】

上記２例とも文中の文字の種類が少ない例を示した。これは計算例として示すために、非常に短い文から確率計算をしたためである。日本語のように文字種が多い場合は、さらに多くの学習用（確率計算用）の文が必要である。

【００４２】

新聞データ約１千万文字からなる文書（文の集合）で文字間の接続確率を計算した例について、その一部を示したものが図７である。この計算結果を使い、文字列「利用者の減少と反比例するように」を計算した結果が図８である。図８では閾値δ＝０．０７で分割点を決定している。

【００４３】

本発明の第一の実施の形態においては、分割処理対象文書自身から文字間接続確率を計算し、その確率を使って同じ分割処理対象文書を分割した。この方法は対象文書に出現する文字の組み合わせの確率全てを計算できるという点で合理的である。

【００４４】

なお、本発明は、処理対象文書自身だけから文字間接続確率を計算するというものに限定されるものではない。まとまった文書から文字間接続確率を計算して

おき、それを使って別の文書を分割することも可能である。これは漸増的な文書データベースにおいて有効である。この場合は分割対象文書に出現する文字の組み合わせが、確率を計算（学習）した文書に出現していない可能性も否定できないが、これらはＮグラム平滑化の問題として（参考文献：「単語と辞書」松本祐治　他著、岩波書店　１９９７年）などに記述されている方式で対応できる。

【００４５】

以上のように、本発明の第１の実施の形態では、ステップ１０１で入力された文書からステップ１０２で文字間の接続確率を計算し、この接続確率を使ってステップ１０３で該文書のそれぞれの文字の接続確率を調べることで単語分割を行い、ステップ１０４で結果を出力することで、辞書を使わない単語分割を行うことが可能になり、その実用的効果は大きい。

【００４６】

（実施の形態２）

本発明の第２の実施の形態の文字列分割装置の構成図は本発明の第１の実施の形態の図２と同じものである。また動作の概要は、本発明第１の実施の形態の図２と同じであるが、計算方式として別のものを用い、よって図１のステップ１０２、およびステップ１０３の手順が変更されるので、その詳細を説明する。

【００４７】

本発明の第１の実施の形態の説明では、文字列接続確率の計算にはＮグラムを用いることで、文字列Ｃi-N+1...Ｃi-1・が出現したという条件のもとで文字Ｃi・が出現する確率を使った。（式（２）参照）例えば文中に出現する「ａｂｃｄｅｆ」の「ａｂｃ」と「ｄｅｆ」の間の接続確率を計算するために、「ａｂｃ」という文字列が出現した場合に次が「ｄ」である確率を使ったのである。これは既存の技術であるＮグラム方式を転用して用いたからである。Ｎグラムはもともと、単語の接続の確からしさや文字の接続の確からしさを計算し、文全体として正しいかを判断するためのものである。または、いままでに出現した単語列や文字列から次の単語や文字を予想するものである。よって本来は、

【００４８】

【数４】

$$\prod_{i=1}^{m} P(w_i|w_1 w_2 \ldots w_{i-1})$$

という確率式で計算するものを

【００４９】

【数５】

$$\prod_{i=1}^{m} P(w_i|w_{i-N+1} \ldots w_{i-1})$$

と近似した場合の、product･記号Πの中の項であった。つまり全ての項を掛け合わせる形で使うことが前提だったので、文字列Ｃi-N+1...Ｃi-1･が出現したという条件のもとで文字Ｃi･が出現する確率というように、条件の部分が複数文字･･（文字列）で、その後に特定の１つの文字が出現する確率で扱えたのである。

【００５０】

しかし本発明は文字間接続確率を、単語内での文字接続か、単語間の文字接続かを判別するために用いる。よって本発明の第２の実施の形態では、文字Ｃi-1･と文字Ｃi･の接続確率を、ある文字列が出現したという条件でのある１文字の出現確率で表現するのではなく、ある文字列が出現したという条件でのある文字列の出現確率を計算するという方式にする。

【００５１】

形式的に表現するなら、長さｎ個の文字列Ｃi-n...Ｃi-1･が出現したという条件のもとで長さ１個の文字列Ｃi･が出現する確率を計算するよりも、長さｎ個の文字列Ｃi-n...　Ｃi-1･が出現したという条件のもとで長さｍ個の文字列Ｃi...Ｃi+m-1･が出現する確率を計算する。

【００５２】

この確率を式（２）と同様に書くならば、次式（４）のようになる。

【００５３】

【数６】

$$P(\underbrace{C_i \cdots C_{i+m-1}}_{m\text{個}} | \underbrace{C_{i-n} \cdots C_{i-1}}_{n\text{個}}) \qquad (4)$$

【００５４】

例えば文中に出現する「ａｂｃｄｅｆ」の「ａｂｃ」と「ｄｅｆ」の間の接続確率を計算するために、「ａｂｃ」という文字列が出現した場合に次が「ｄｅ‥ｆ」である確率を使うのである。これはｎ＝３、ｍ＝３の例である。式（４）はｍ＝１の場合が本発明の第１の実施の形態に相当する。

【００５５】

また第１の実施の形態が、文の先頭側にある文字列から次の文字列への接続確率を求める、つまり前から後へ進む順方向の確率の計算と考える時、式（４）のｎ＝１、ｍ＞１という条件は、後から前への接続確率の計算に近似でき、逆方向の確率の計算に相当する。例えば文中に出現する「ａｂｃｄｅｆ」の「ａｂｃ」と「ｄｅｆ」の間の接続確率を計算するために、ｎ＝１、ｍ＝３なら、文字「‥ｃ」の後に文字列「ｄｅｆ」が出現する確率ということになるが、文字列「ｄｅｆ」が出現した場合にその前が「ｃ」である確率に近似できる。これは逆方向の文字間接続確率の計算に相当する。ところが式（４）の計算のためには、ｎ＋ｍグラムの統計を取る必要がある。しかし、ｎ≧２かつｍ≧２とすると、４グラム（またはそれ以上）の文字組の統計計算が必要になり、非常に大きな記憶空間を必要とする。

【００５６】

そこで本発明の第２の実施の形態では、式（４）の計算を次式（５）で近似する方式を提案する。

【００５７】

【数７】

$$P(C_i|C_{i-n}\ldots C_{i-1}) \times P(C_{i-1}|C_i\ldots C_{i+m-1}) \qquad (5)$$

【００５８】

式（５）は、ｎ個の文字列が出現した後に特定の文字が出現する順方向の確率である第１項と、ｍ個の文字列が出現する前に特定の文字が出現する逆方向の確率である第２項の積である。項と文字列の関係を図１４に示す。例えば文中に出現する「ａｂｃｄｅｆ」の「ａｂｃ」と「ｄｅｆ」の間の接続確率を計算するために、「ａｂｃ」の後に「ｄ」が出現する確率（順方向の第１項）と、「ｄｅ‥ｆ」が出現した時にその前が「c･」である確率（逆方向の第２項）の積を取ることを意味する。

【００５９】

式（５）の確率値は、第１項はｎ＋１グラム、第２項はｍ＋１グラムの統計を取ればよく、式（６）により計算（推定）できる。

【００６０】

【数８】

$$\underbrace{\frac{Count(C_{i-n}\ldots C_i)}{Count(C_{i-n}\ldots C_{i-1})}}_{\text{第１項}} \times \underbrace{\frac{Count(C_{i-1}\ldots C_{i+m-1})}{Count(C_i\ldots C_{i+m-1})}}_{\text{第２項}} \qquad (6)$$

【００６１】

式（６）を計算し、文字ｎ＋１個組、および文字ｍ＋１個組とともに計算結果を、図２の確率テーブル格納手段２０４に記録することが、本発明の第２の実施の形態における図１のステップ１０２に相当する。よって、図２の確率テーブル格納手段２０４は、文字ｎ＋１個組用と文字ｍ＋１個組用の２つのテーブルを持つことになる。ｎ≠ｍの場合、この計算手順は、例えば図９に示す手順で実現で

きる。

ステップ９０１：文書を構成する文ごとに前後に文頭文末を表現する特殊記号を、文頭にはｎ－２個、文末にはｍ－２個付与する。本発明の第２の実施の形態では、後から前への接続確率も計算するため、文の後にも特殊記号を付与する必要がある。

ステップ９０２：ｎグラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｎ個組について、それが何回出現しているかを調べた表を作成する。

ステップ９０３：ｎ＋１グラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｎ＋１個組について、それが何回出現しているかを調べた表を作成する。

ステップ９０４：ｎ＋１グラム統計の夫々の文字ｎ＋１個組文字列について、その出現回数をＸとする。同文字ｎ＋１個組文字列について、その１番目からｎ番目の文字列の出現回数をステップ９０２で作成したｎグラム統計から調べ、これをＹとする。Ｘ／Ｙにより式（６）の第１項の値を計算し、計算結果を確率テーブル格納手段２０４の文字ｎ＋１個組（第１項の確率値）の部分に記録する。

ステップ９０５：ｍグラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｍ個組について、それが何回出現しているかを調べた表を作成する。

ステップ９０６：ｍ＋１グラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｍ＋１個組について、それが何回出現しているかを調べた表を作成する。

ステップ９０７：ｍ＋１グラム統計の夫々の文字ｍ＋１個組文字列について、その出現回数をＸとする。同文字ｍ＋１個組文字列について、その２番目からｍ＋１番目の文字列の出現回数をステップ９０５で作成したｍグラム統計から調べ、これをＹとする。Ｘ／Ｙにより式（６）の第２項の値を計算し、計算結果を確率テーブル格納手段２０４の文字ｍ個組（第２項の確率値）の部分に記録する。

【００６２】

以上がｎ≠ｍの場合である。ｎ＝ｍの場合、図２の確率テーブル格納手段２０４は、文字ｎ個組用だけを用意すればよく、その構造は図１２（ｄ）のようにｎ文字組とその第１項の確率、第２項の確率を記録するものとなる。そしてｎ＝ｍ

の場合、計算手順は例えば図１０に示すように図９よりも簡略化できる。

ステップ１００１：文書を構成する文ごとに前後に文頭文末を表現する特殊記号をｎ－２個ずつ付与する。

ステップ１００２：ｎグラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｎ個組について、それが何回出現しているかを調べた表を作成する。

ステップ１００３：ｎ＋１グラム統計を作成する。すなわち、対象文書の中に出現した全ての文字ｎ＋１個組について、それが何回出現しているかを調べた表を作成する。

ステップ１００４：ｎ＋１グラム統計の夫々の文字ｎ＋１個組文字列について、その出現回数をＸとする。同文字Ｎ個組文字列について、その１番目からｎ番目の文字列の出現回数をステップ１００２で作成したｎグラム統計から調べ、これをＹとする。Ｘ／Ｙにより式（６）の第１項の値を計算し、計算結果を確率テーブル格納手段２０４の確率値第１項の部分に記録する。

ステップ１００５：Ｎグラム統計の夫々の文字ｎ＋１個組文字列について、その出現回数をＸとする。同文字Ｎ個組文字列について、その２番目からｎ＋１番目の文字列の出現回数をステップ１００２で作成したｎグラム統計から調べ、これをＹとする。Ｘ／Ｙにより式（６）の第２項の値を計算し、計算結果を確率テーブル格納手段２０４の確率値第２項の部分に記録する。

【００６３】

以上により式（６）の値が計算できる下地が整ったが、まだ式（６）の値そのものは求めてなく、実際の値は次の分割過程で計算する。以下では図１のステップ１０３の詳細を説明する。ステップ１０３はステップ１０２で計算した文字間接続確率の表を使って、処理対象の文を構成する文字のそれぞれの部分の接続確率を計算し、分割をする過程である。その計算手順を、ｎ≠ｍの場合について図１１に示す。

ステップ１１０１：文書から文を一つ選択する。

ステップ１１０２：図１０のステップ１００１と同様に、文の前後に文頭文末を表現する特殊記号を、文頭にはｎ－２個、文末にはｍ－２個付与する。

ステップ１１０３：ポインタを文の前に付けた特殊記号の一文字目に移動する。
ステップ１１０４：ポインタ位置から始まるｎ＋１文字について、確率テーブル格納手段２０４の確率値第１項から検索し、それをポインタ位置を１文字目とした時のｎ文字目とｎ＋１文字目の間の確率値第１項として記録する。ただし文字と文頭文末特殊記号の間の接続確率は０とする。
ステップ１１０５：同様にポインタ位置から始まるｍ＋１文字について、確率テーブル格納手段２０４の確率値第２項から検索し、それをポインタ位置を１文字目とした時の１文字目と２文字目の間の確率値第２項として記録する。ただし文字と文頭文末特殊記号の間の接続確率は０とする。
ステップ１１０６：ポインタを一文字進める。
ステップ１１０７：ポインタが文末の文字を指していたら、文は終了したものとして、ステップ１１０８へ進み、そうでなければステップ１１０４へジャンプする。
ステップ１１０８：各文字間について、各文字間に記録された確率値第１項と確率値第２項の積を取り、式（６）の値を計算する。それがあらかじめ決められた閾値δ未満だったら、そこで分割を行う。もしその確率が閾値δ以上だったら、そこは単語分割点ではないので分割を行わない。
ステップ１１０９：文書から次の文を選択する。
ステップ１１１０：次の文が無ければ終了。そうでなければステップ１１０２へ進む。以上により分割点を発見する。ｎ＝ｍの場合も、同様である。

【００６４】

以下、具体的計算例を示す。全文書として文字列「仕事は仕事」だけが与えられた場合を例とし、ここから文字間接続確率をｎ＝ｍ＝２の（ｎ＋１）グラム‥（３グラム）で計算する。まずステップ１００１として、文（文字列）の前後に文頭文末を表現する特殊記号をｎ－２（３－２＝１）個ずつ付与する。この様子を図１２（ａ）に示す。特殊記号としてここでは＃を付けているが、実際には文書に現れない記号を付けるものとする。次にステップ１００２として、２グラムの統計、すなわち文字２個組の出現回数を調べる。その結果が図１２（ｂ）のようになる。同様にステップ１００３として３グラムの、文字３個組の出現回数を

調べ、図１２（ｃ）を得る。またステップ１００４として、図１２（ｂ）と図１２（ｃ）から、文字３個組についての式（６）の第１項の値を計算し、図１２‥（ｄ）の第１項の部分を得る。また、ステップ１００５として、図１２（ｂ）と図１２（ｃ）から、文字３個組についての式（６）の第２項の値を計算し、図１２（ｄ）の第２項の部分を得る。

【００６５】

注意しなければいけないのは、図１２（ｄ）は同じ箇所の文字間接続確率を記録したものではない。図１２（ｄ）の表の第２行目「仕事は」について、第１項の部分に記録された確率は「仕事」と「は」の接続確率第１項であり、第２項の部分に記録された確率は「仕」と「事は」の接続確率第２項である。以上により確率値のテーブルができたので、次に図１１の処理に進む。

【００６６】

ステップ１１０１として「仕事は仕事」が選択され、ステップ１１０２で前後に特殊文字が付けられることで図１２（ａ）と同じ状態になる。ステップ１１０３からステップ１１０５で図１２（ｅ）を得る。第２項は文頭特殊記号と文字との間なので０とする。同様にステップ１１０３から１１０６までの繰り返しで図１２‥（ｆ）を得る。ステップ１１０８により各文字間の接続確率が計算され、図１２（ｆ）の接続確率の部分を得る。あらかじめ決められた閾値δ（ここでは閾値δ＝０．６とする）未満の部分で分割を行うと、図１２（ｆ）に示す通り、「仕事／は／仕事」という分割結果を得る。

【００６７】

なお閾値δはあらかじめ決められたものとして扱ってきたが、確率の値を計算した後、単語分割結果が望むべく平均単語長を満すように動的に決めてもよい。すなわち、図１３に示すように、閾値δが大きければ平均単語長は長くなり、閾値δを小さくすれば平均単語長は短かくなる。分割結果で調整しながらδを文書ごとに決めるようにすれば、適切な値が取れるようになる。

【００６８】

また、閾値δは一率として扱ってきたが、何らかの基準により複数設定してもよい。日本語の場合は本来、平仮名部分の平均単語長は漢字部分の平均単語長よ

りも短い傾向にある。これは平仮名が助詞などの一文字の単語を含むからである。その一方で片仮名部分は外国語の発音を表記したものが多いことから平均単語長は長い。よって閾値δを文字種（漢字・平仮名・片仮名）により複数設定してもよい。

【００６９】

また、日本語の場合、単語分割点は文字種（漢字・平仮名・片仮名）の変化点にあることが多い。よって文字種の変化点の閾値δを他の部分より下げるなどしてより適切な値に調整してもよい。

【００７０】

なお、本発明の実施の形態では文頭や文末は必ず単語分割点であるとして説明してきたが、この他に句読点の前後、カッコや記号の前後も単語の分割点とみなしてよく、その部分の確率計算を省略することが可能である。あるいはＮグラム統計作成において、句読点や記号の前後も文の切れ目として計算してよい。すなわち、本発明第２の実施例で用いた「仕事は仕事」の３グラムは、文頭文末特殊記号を付与することで「＃仕事は仕事＃」という文字列を作って計算した。これが「仕事は、仕事」だった場合、文は２つであるとみなし、「＃仕事は＃」と‥「＃仕事＃」の２つの文について計算するようにしてもよい。また、式（５）で第１項と第２項の積を取ることで確率を計算したが、第１項と第２項の和、あるいは加重平均などにより確率を計算してもよい。

【００７１】

以上のように、本発明の第２の実施の形態では、文字ｎ個の後に文字ｍ個が続く確率を近似式（６）で計算することで、より正確に、辞書を使わない単語分割を行うことが可能になり、その実用的効果は大きい。

【００７２】

（実施の形態３）

第３の実施の形態の文字列分割装置は、あらかじめ作成されている単語辞書を持ち、この辞書を利用しながら文字列を単語単位に分割するが、その過程で本発明の第１、または第２の実施の形態で用いた文字間接続確率を利用する。

【００７３】

まず原理を説明する。

【００７４】

「小田中学校」という文字列を分割することを考える。単語辞書に「学校」・・「小田」「小田中」「中学校」という４つの単語があったとすると、分割候補は図１７（ａ）のように「小田／中学校」と「小田中／学校」の２つが考えられる。辞書だけの情報ではこれら２つの候補から１つを選択するのは難しい。

【００７５】

そこで本発明では「文字間接続確率が小さい部分は単語分割点として尤もらしい」として、複数の単語分割解の候補のうち接続確率が小さい所で分割されているものを選択する。図１７（ａ）の例では文字間接続確率が図１７（ｂ）のように計算されているとすると、単語の分割点の部分に注目して文字間接続確率Ｐ2･とＰ3･を比較し、Ｐ2･の方が小さいことから候補（１）の「小田／中学校」を正解として選択する。

【００７６】

この原理は長い文字列に対しても同様に適用できる。「大阪市立山田中学校」という文字列を分割することを考える。単語辞書に「学校」「山田」「市立」・・「大阪」「大阪市」「中学」「中学校」「田中」「立山」といった単語が存在する時、分割解の候補は図１８（ａ）に示すように「大阪／市立／山田／中学校」と「大阪市／立山／田中／学校」の２つの候補が考えられる。この２つのどちらを正解とするかについても、辞書だけの情報から選択するのは難しい。そこで先の例と同様にこの解の候補の選択にも文字間接続確率を用いて、接続確率が小さい所で分割されているものを選択する。

【００７７】

長い文字列の場合は、一つの分割解候補の中に複数の分割点があることから、これら分割点の確率の和や積から分割解の候補ごとにスコアを求め、スコアの比較から候補を選択する。図１８（ａ）の例では、文字間の接続確率が図１８（ｂ）のように計算されているとすると、各候補のスコアを分割点での文字間接続確率の和として計算することで図１８（ｃ）を得る。よってスコアの値の小さい候補（１）の「大阪／市立／山田／中学校」を正解とする。

【００７８】

上記の原理にのっとり、本発明の第３の実施の形態について、図面を用いてその処理手順を説明する。図１５は本発明の第３の実施の形態を示した構成図の一例である。第３の実施の形態の文字列分割装置は、処理対象文書を電子化された状態で入力するための文書入力手段２０１と、入力した文書を蓄えておく文書蓄積手段２０２と、文書から文字間接続確率を計算するための文字間接続確率計算手段２０３と、計算した確率を記録しておくための確率テーブル格納手段２０４と、あらかじめ作成されている単語辞書を記憶するための単語辞書記憶手段２０７と、単語辞書の内容を用いて分割解の候補を作成するための分割解候補作成手段２０８と、解の候補の中から文字間接続確率を用いて解を選択するための解選択手段２０９と、処理結果の文書を出力する出力手段２０６を備えている。

【００７９】

処理全体の流れを図１６で示す。

ステップ１６０１：文書入力手段２０１から入力されたデータは、まず文書蓄積手段２０２に蓄えられる。

ステップ１６０２：このデータから、文字間の接続確率を文字間接続確率計算手段２０３が計算し、計算結果を確率テーブル格納手段２０４に蓄える。計算方法の詳細は第１の実施の形態、または第２の実施の形態の方式に依る。

ステップ１６０３：文書蓄積手段２０２に蓄えられたデータについて、単語辞書記憶手段２０７に蓄えられた単語辞書情報から分割解候補作成手段２０８が分割解の候補を作成し、この分割候補の中から確率テーブル格納手段２０４に蓄えられた確率値を用いることで解選択手段２０９が候補を選択し、その結果として文字列が単語単位に分割され（詳細は後述）、

ステップ１６０４：分割された文を文書出力手段２０６から出力する。

【００８０】

以上のように本発明における文字列分割装置は、処理対象文書から文字間接続確率を計算し、計算した確率と単語辞書の情報を使って対象文書を単語単位へ分割することを特徴とする。以下、図１６のステップ１６０３の詳細について、図を用いて説明する。図１６のステップ１６０３は、図１９で示す処理を行う。

ステップ１９０１：分割しようとする文字列の最初の文字から順に、そこから始まる単語が単語辞書記憶手段２０７の中にあるかを調べ、あればそれらを羅列する。文字列「大阪市立山田中学校」を例ととすると、この状態は図２０に相当する。

ステップ１９０２：単語を結ぶことで文の最後まで到達するものを解の候補とする。そして夫々の解の候補のスコアを計算する。スコアは、夫々の解の候補における各単語分割点の文字間接続確率の和とする。文字列「大阪市立山田中学校」を例とすると、解の候補は図１８（ａ）、文字間接続確率が図１８（ｂ）の通りの時、夫々の解の候補のスコアは図１８（ｃ）となる。

ステップ１９０３：最も小さいスコアを持つものを解として選択する。図１８・・（ｃ）では候補（１）を選択する。

【００８１】

以上により単語分割を得る。文字間接続確率は必ず０以上であり、ステップ１９０２で文字間接続確率の和を取ったことから、文字列の分割について「中／学校」と「中学校」のように一つの単語になるか複数に分割されるかで複数の候補があった場合は、分割点が少ない方が必ず選択される。ステップ１９０２では分割解の候補のスコアとして文字間接続確率の和を取り、そのスコアが最小になるものを解とした。これを式で表現すると式（７）のようになる。すなわち、単語分割位置の集合をＳとする時、Ｓに含まれる全ての位置ｉにおける文字間接続確率Ｐｉ･の和が最小となるような集合Ｓを選択するという意味である。

【００８２】

【数９】

$$\arg\min_{S} \sum_{i \in S} Pi \qquad (7)$$

【００８３】

本発明はスコアの計算は、確率の和に限定されるものではない。例えばスコア

の計算として文字間接続確率の積を取ってもよい。これを式（７）と同様に記述するなら、次の式（８）のようになる。

【００８４】

【数１０】

$$\arg\min_{S} \prod_{i \in S} Pi \qquad (8)$$

【００８５】

確率の積を計算する場合は、同じ効果を得るものとして対数を取ってもよい。対数を取ることにより、積の計算が対数の和へ換算することができる。これは式（９）式（１０）のように書くことができる。

【００８６】

【数１１】

$$\arg\min_{S} \log\left(\prod_{i \in S} Pi\right) \qquad (9)$$

$$= \arg\min_{S} \sum_{i \in S} \log Pi \qquad (10)$$

【００８７】

なお、本発明は文字間接続確率を導入して単語分割の解の候補のスコア計算と選択の方法を提案とするものであり、解候補を得るために単語を結んでスコアを計算していく手順については、本実施の形態の手順に特定するものではない。一般にステップ１９０２のような状況で単語を結んで、図１８（ａ）のような解の候補を作成し、同時にスコア（あるいはコスト）を計算する場合、計算手法としては動的計画法などの探索手法が提案されており、この動的計画法のアルゴリズムを用いてもよい。

【００８８】

以上のように、本発明の第３の実施の形態においては、単語分割処理対象文書自身から文字間接続確率を計算しておき、分割解の候補の作成には単語辞書を用いるが、複数の解候補から一つを選択するには文字間接続確率が最小になるものを選択する。よって分割候補選択のための知識の学習に、あらかじめ人手で作成された分割正解文書を大量に用意する必要がないことからコストを抑えることが可能であり、分割対象の文書から確率の形で知識を自動学習することが可能で、文書分野に適合した学習ができる点で合理的であり、その実用的効果は大きい。

【００８９】

（実施の形態４）

以下、本発明の第４の実施の形態について、図面を用いて説明する。図２１は本発明の第４の実施の形態を示した構成図の一例である。

【００９０】

第４の実施の形態の文字列分割装置は、処理対象文書を電子化された状態で入力するための文書入力手段２０１と、入力した文書を蓄えておく文書蓄積手段２０２と、文書から文字間接続確率を計算するための文字間接続確率計算手段２０３と、計算した確率を記録しておくための確率テーブル格納手段２０４と、あらかじめ作成されている単語辞書を記憶するための単語辞書記憶手段２０７と、未知語候補を特定する未知語特定手段２１０と、単語辞書の内容と未知語特定手段の結果を用いて分割解の候補を作成するための分割解候補作成手段２０８と、解の候補の中から文字間接続確率を用いて解を選択するための解選択手段２０９と、処理結果の文書を出力する出力手段２０６を備えている。

【００９１】

第４の実施の形態の文字列分割装置の処理全体の流れを図２２で示す。

ステップ２２０１：文書入力手段２０１から入力されたデータは、まず文書蓄積手段２０２に蓄えられる。

ステップ２２０２：このデータから、文字間の接続確率を文字間接続確率計算手段２０３が計算し、計算結果を確率テーブル格納手段２０４に蓄える。計算方法の詳細は第１の実施の形態、または第２の実施の形態の方式に依る。

ステップ２２０３：文書蓄積手段２０２に蓄えられたデータについて、単語辞書記憶手段２０７に蓄えられた単語辞書情報と未知語特定手段２１０が特定した未知語候補を使い、分割解候補作成手段２０８が分割解の候補を作成し、この分割候補の中から確率テーブル格納手段２０４に蓄えられた確率値を用いることで解選択手段２０９が候補を選択し、その結果として文字列が単語単位に分割される。

ステップ２２０４：分割された文を、出力手段２０６から出力する。

【００９２】

以上のように本発明における文字列分割装置は、処理対象文書から文字間接続確率を計算し、計算した確率と単語辞書の情報、および未知語推定手段が推定した未知語候補を使って対象文書を単語単位へ分割することを特徴とする。以下、図２２のステップ２２０３の詳細について、図と例を用いて説明する。図２２のステップ２２０３は、図２３で示す処理を行う。例として文字列「大阪市立山田中学校」を用いる。単語辞書記憶手段２０７には「学校」「市立」「大阪」「大阪市」「中学」「中学校」「田中」「立山」という単語があるが「山田」は単語として登録されていないものとする。（図２４）

ステップ２３０１：分割しようとする文字列の最初の文字から順に、そこから始まる単語が単語辞書記憶手段２０７の中にあるかを調べ、あればそれらを羅列する。文字列「大阪市立山田中学校」を例とすると、この状態は図２５（ａ）に相当する。図２５（ａ）の状態では「山田」はまだ単語として認められていない。

ステップ２３０２：ある文字位置ｉで、その直前の文字位置ｉ－１で終了する単語が存在するのに文字位置ｉから開始される単語が存在しない時、文字位置ｉから始まる長さｎ文字以上ｍ文字以下の全ての文字列を未知語として羅列する。図２５（ａ）の例では５番目の文字「山」の直前で終わる単語「市立」が存在するのに、「山」から始まる単語が存在しない。そこで「山」から始まる長さｎ～ｍ文字の文字列を未知語として先の羅列に加える。ｎとして２、ｍとして３を取った場合、「山田」と「山田中」の２つが未知語として推定される。この様子が図２５（ｂ）に相当する。

ステップ２３０３：単語を結ぶことで文の最後まで到達するものを解の候補とす

る。図２５（ｂ）の中から単語を結ぶことで最後まで到達するものは、図２６に示すグラフのようになることから、図２７に示す３つが解の候補となる。そして夫々の解の候補のスコアを計算する。スコアは、夫々の解の候補における各単語分割点の文字間接続確率の和とする。図２７の分割解候補の夫々のスコアは、文字間接続確率が図１８（ｂ）の通りとすると、図２７の右端に示す通りである。ステップ２３０４：最も小さいスコアを持つものを解として選択する。図２７では候補（１）が選択される。

【００９３】

以上により単語分割を得る。ステップ２３０３では分割解の候補のスコアとして文字間接続確率の和を取ったが、本発明はスコアの計算はこれに限定されるものではない。スコアの計算方法は式（７）の他に、式（８）式（９）式（１０）などが考えられる。上記は単語分割の解の候補について、分割部分の文字間接続確率値だけでスコアを計算したが、本発明では分割しない部分についても特定の値を式に加えることを提案する。

【００９４】

解候補のうち、単語分割点については文字間接続確率を、単語分割点以外の文字間には定数を与え、それらから候補のスコアを計算する。より形式的に説明するなら、以下のようになる。文字の位置についての集合をＮ、単語分割する文字位置の集合をＳとする時（Ｓ⊆Ｎ）、各文字位置ｉについて値Ｑｉ･を次のように定める。すなわち、Ｓに含まれる位置ｉについては文字間接続確率を、Ｓに含まれない位置ｉについては定数値Ｔｈを与え（式（１２））、解の候補についてこの値の和を取ったものをスコアとし、スコアが最小になるものを解として選択する（式（１１）あるいは式（１３））。定数値Ｔｈは本発明の第１の実施の形態における閾値δに相当する。

【００９５】

【数１２】

$$\arg\min_{S\subseteq N}\sum_{i\in N}Qi \qquad (11)$$

$$Q_i=\begin{cases}P_i, & i\in S\\ Th, & i\notin S\end{cases} \qquad (12)$$

【００９６】

【数１３】

$$\arg\min_{S\subseteq N}\left(\sum_{i\in S}Pi+\sum_{i\notin S}Th\right) \qquad (13)$$

【００９７】

例として文字列「新宿泊棟」（新しい宿泊用建物の意味）の分割過程を見る。「新宿泊棟」の分割候補として図２８（ａ）のように「新／宿泊／棟」と「新宿／泊棟」の２つがあったとする。後者は「泊棟」が未知語推定されたものとする。文字間接続確率は図２６（ｂ）のように与えられているものとする。この例の場合、先に示した式（７）だけで計算すると、図２８（ａ）の候補（１）はスコアが０．０４４に、候補（２）はスコアが０．０４０になり、候補（２）が誤って選択されてしまう。

【００９８】

これに対して先に示した式（１１）の計算方法を採用すると、Ｔｈとして０．０３を与えた時、図２８（ｃ）のように候補（１）は「宿泊」の「宿」と「泊」の間は単語に切れないのでこの位置に定数Ｔｈを与えてスコアを計算することで、候補（１）のスコアとして０．０７４を得る。同様に候補（２）は「新」と「‥宿」の間、および「泊」と「棟」の間に定数Ｔｈを与えて計算することで、候補（２）のスコアとして０．１００を得る。これらを比較することで候補（１）

を選択することができる。

【００９９】

式（７）を用いた分割が分割数が最小になるものを選択する傾向にあるのに対して、式（１１）を用いた計算は、より細かく分割した解候補の選択をも可能とする。つまり、「衆議院議員」のような複合語が辞書にあったとしても、それを「衆議院」と「議員」という短い単位の単語に分解できる。単語分割装置の使用目的によっては、文書によって分割の粒度を制御する要求があるが、Ｔｈをパラメータとすることでこれを実現可能とする。

【０１００】

式（１１）は、スコアに和を取る場合の式（７）の計算に閾値に相当する値を導入するものであるが、式（８）の積の場合にも同様に閾値を導入することで、分割粒度を制御することが可能になる。

【０１０１】

また、これまでに述べたスコア計算方法は単語分割位置についての文字間接続確率と、単語分割されない位置についての定数値のみを用いたが、本発明では夫々の単語に何らかのスコアを与えることも提案する。単語へのスコアの与え方として、それが単語辞書記憶手段２０７に記憶されていたものであれば定数Ｕを、未知語推定手段２１０が推定した未知語であれば定数Ｖを与えるものとし、候補中の単語の集合をW、単語辞書記憶手段２０７に記憶されている単語の集合をＤとするとき、式（１１）を拡張したスコアは次の式により記述できる。

【０１０２】

【数１４】

$$\arg \min_{S \subseteq N, W} \left(\sum_{i \in N} Qi + \sum_{j \in W} R_j \right) \qquad (14)$$

$$Q_i = \begin{cases} P_i, & i \in S \\ Th, & i \notin S \end{cases} \qquad (15)$$

$$R_j = \begin{cases} U & j \in D \\ V & j \notin D \end{cases} \qquad (16)$$

$$U < V \qquad (17)$$

【０１０３】

この時、Ｕ＜Ｖとすることで未知語よりも辞書に載っている単語を優先して選択することを可能とする。よって、単語分割候補の中に含まれている未知語が少ないものが解として選択されるという効果を生む。

【０１０４】

未知語推定手段２１０の導入と、定数Ｔｈの導入、および単語のスコアの導入により、未知語推定をしながら推定した未知語を選択するか否か判断しながら単語分割をすることになる。未知語推定手段は図２３のステップ２３０２で辞書に無い長さｎ～ｍ文字の任意の文字列を未知語候補とした。この未知語を文字間接続確率によって選択することは、未知語の部分は本発明の第１または第２の実施の形態のように確率値だけで分割することと等価になる。よって、辞書による単語分割と確率による単語分割の統合が可能となる。

【０１０５】

従来の技術では未知語の推定には、「漢字と平仮名の境界は単語の分割点になりやすい」といった経験による知識を用いていたが、本発明においては図２３のステップ２３０２において条件範囲内の全ての文字列を未知語として扱う。しかし文字間接続確率の計算をすることでその中から正解を選択することができる。これは例えば漢字と平仮名にまたがった未知語も推定できることを意味している

。

【０１０６】

以上のように、本発明の第４の実施の形態においては、分割処理対象文書自身から文字間接続確率を計算しておき、単語辞書と未知語推定手段を使って分割解の候補を作成するが、複数の解候補から一つを選択するには文字間接続確率が最小になるものを選択する。辞書にない単語については未知語を推定するが、未知語を選択するかどうかは確率値（あるいは計算されたスコア）に従うことで、未知語周辺は確率値により分割を行う。よって分割候補選択のための知識の学習に、あらかじめ人手で作成された分割正解文書を大量に用意する必要がないことからコストを抑えることが可能であり、分割対象の文書から確率の形で知識を自動学習することが可能で、文書分野に適合した学習ができる点で合理的であり、その実用的効果は大きい。さらに辞書にない未知語は確率値から分割することも可能であり、その実用的効果は大きい。

【０１０７】

【発明の効果】

以上のように、本発明では、処理対象文書の中から文字間接続確率を計算し、その確率を処理対象文書にあてはめることで単語分割できる場所を発見して分割するものであり、これにより、辞書を使わずにテキストを単語に分割するという効果を奏するものである。よって本発明は単語の分割のために辞書を用意する必要がないことから、日々生まれ続ける新しい語や語法のために、辞書を整備したり各種パラメータを整備する必要もない。辞書を持たないことから、本方式により実現されたプログラムを格納した記録媒体は非常に小さくて済む。同時にパーソナルコンピュータなどの処理能力に限界のある環境下においても機能する。

【０１０８】

また、辞書を用いた単語分割においても文字間接続確率を用いることで、複数の分割解候補の中から一つを選択することが可能となる。

【０１０９】

さらに、辞書を用いた単語分割において辞書に載っていない単語が出現しても、未知語推定と文字間接続確率を併用することで文字列を単語単位に分割するこ

とが可能となる。

【図面の簡単な説明】

【図１】

本発明の第１の実施の形態における単語分割方式の動作を示すフローチャート

【図２】

本発明の第１の実施の形態における単語分割装置の構成を示すブロック図

【図３】

本発明の第１の実施の形態における文字間接続確率の計算手順を示すフローチャート

【図４】

本発明の第１の実施の形態における単語分割計算例を示す概念図

【図５】

本発明の第１の実施の形態における分割過程の計算手順を示すフローチャート

【図６】

本発明の第１の実施の形態における単語分割計算例を示す概念図

【図７】

本発明の第１の実施の形態における単語分割方式で新聞記事データの文字間接続確率を計算した例の一部を示す数値図

【図８】

本発明の第１の実施の形態における単語分割方式で新聞記事データの分割をした例を示す概念図

【図９】

本発明の第２の実施の形態における文字間接続確率の計算手順について、ｎとｍが違う場合の計算手順を示すフローチャート

【図１０】

本発明の第２の実施の形態における文字間接続確率の計算手順について、ｎとｍが同じ場合の計算手順を示すフローチャート

【図１１】

本発明の第２の実施の形態における分割過程の計算手順を示すフローチャート

【図１２】

本発明の第２の実施の形態における単語分割計算例を示す概念図

【図１３】

本発明における閾値と平均単語長の関係を示す概念図

【図１４】

本発明の第２の実施の形態における式（５）の関係を示す模式図

【図１５】

本発明の第３の実施の形態における単語分割装置の構成を示すブロック図

【図１６】

本発明の第３の実施の形態における単語分割方式の動作を示すフローチャート

【図１７】

本発明の第３の実施の形態における単語分割過程の例を示す概念図

【図１８】

本発明の第３の実施の形態における単語分割過程の例を示す概念図

【図１９】

本発明の第３の実施の形態における単語分割の詳細動作を示すフローチャート

【図２０】

本発明の第３の実施の形態における辞書利用過程を示す概念図

【図２１】

本発明の第４の実施の形態における単語分割装置の構成を示すブロック図

【図２２】

本発明の第４の実施の形態における単語分割方式の動作を示すフローチャート

【図２３】

本発明の第４の実施の形態における単語分割の詳細動作を示すフローチャート

【図２４】

本発明の第４の実施の形態における単語分割の辞書内容の例を示す概念図

【図２５】

本発明の第４の実施の形態における辞書利用と未知語推定の過程を示す概念図

【図２６】

本発明の第４の実施の形態における単語分割過程の例を示す概念図

【図２７】

本発明の第４の実施の形態における分割解の候補の選択過程の例を示す概念図

【図２８】

本発明の第４の実施の形態における分割解の候補の選択方法の例を示す概念図

【符号の説明】

２０１　文書入力手段

２０２　文書データ蓄積手段

２０３　文字間接続確率計算手段

２０４　確率テーブル記憶手段

２０５　文字列分割手段

２０６　文書出力手段

２０７　単語辞書記憶手段

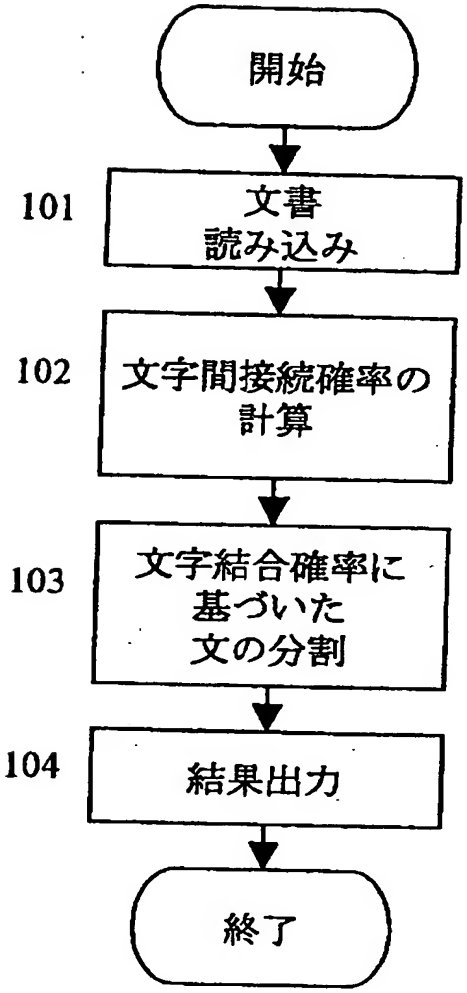
２０８　分割解候補作成手段

２０９　解選択手段

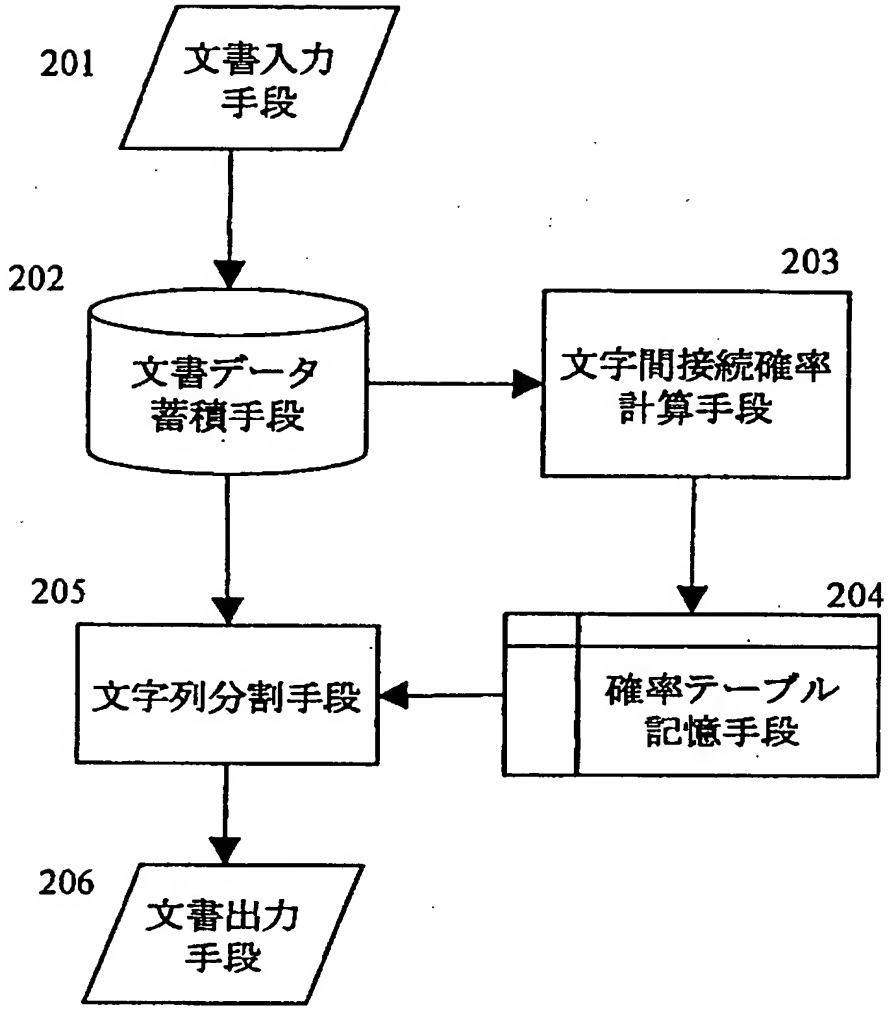
２１０　未知語推定手段

【書類名】　　図面

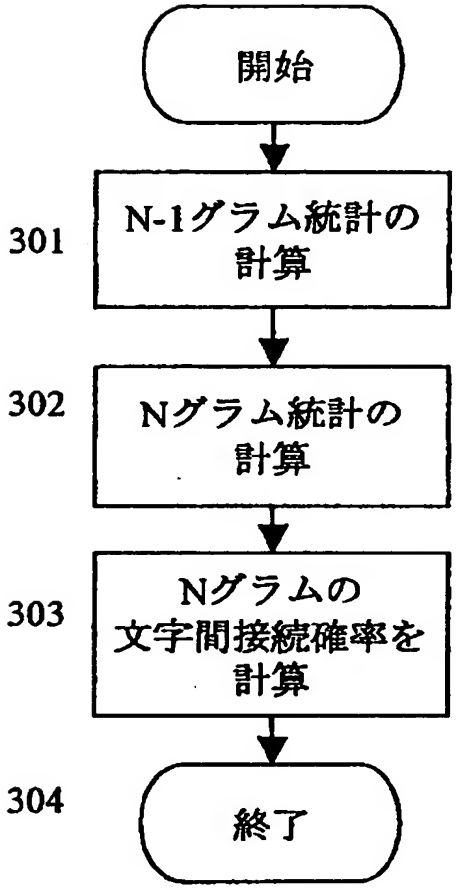
【図１】


開始
101
文書
読み込み
102
文字間接続確率の
計算
103
文字結合確率に
基づいた
文の分割
104
結果出力
終了

【図２】


201
文書入力
手段
202
文書データ
蓄積手段
203
文字間接続確率
計算手段
204
確率テーブル
記憶手段
205
文字列分割手段
206
文書出力
手段

【図３】


開始
301
N-1グラム統計の計算
302
Nグラム統計の計算
303
Nグラムの文字間接続確率を計算
304
終了

【図４】

(a)

処理対象文字列：ａｂａａｂａ

↓

文頭特殊記号付与後の
処理対象文字列　：　＃ａｂａａｂａ

(b)

2グラム出現頻度

文字列	回数
#a	1
ab	2
ba	2
aa	1

(c)

3グラム出現頻度

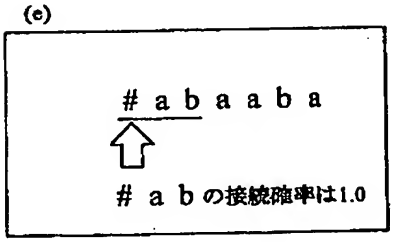
文字列	回数
#ab	1
aba	2
baa	1
aab	1

(d)

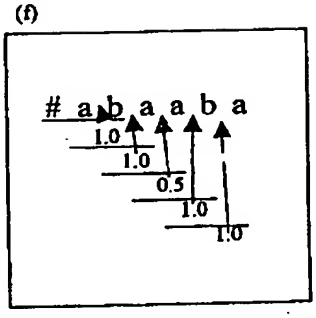
文字間接続確率

文字列	確率	計算方法
#ab	1.0	= Count(#ab)/Count(#a)
aba	1.0	= Count(aba)/Count(ab)
baa	0.5	= Count(baa)/Count(ba)
aab	1.0	= Count(aab)/Count(aa)

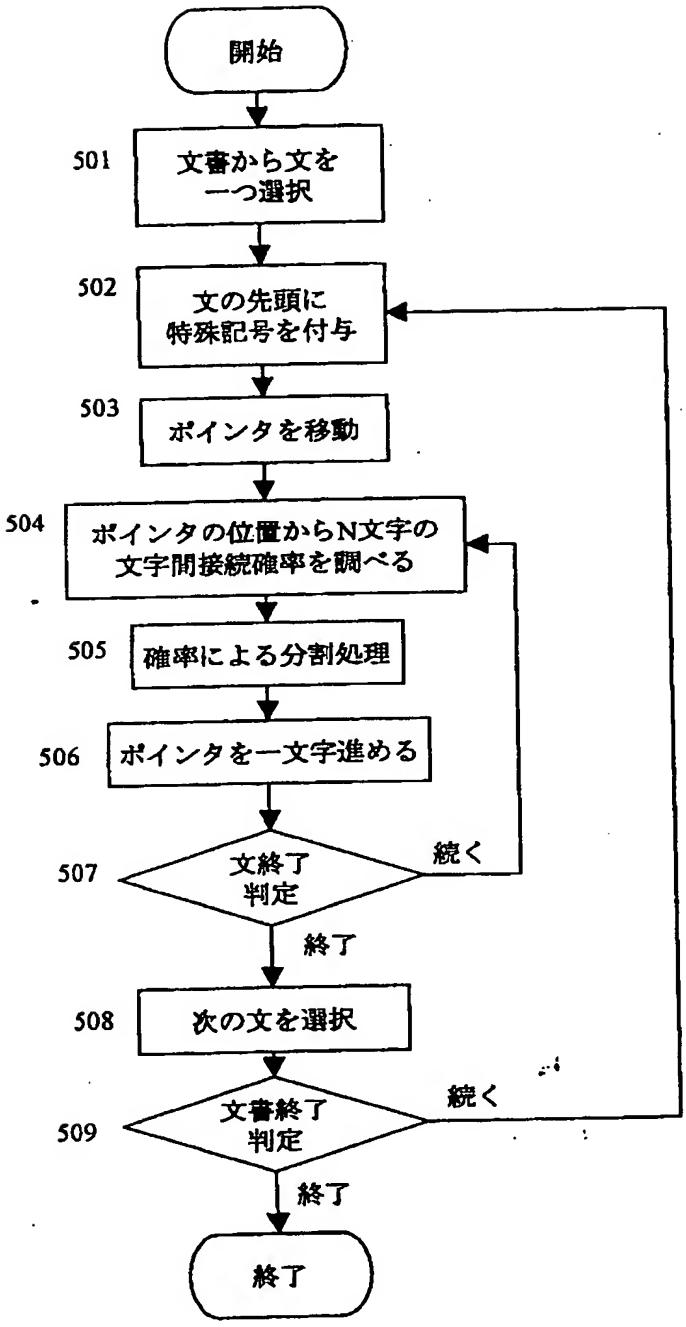
(e)



(f)



【図５】


開始
501
文書から文を一つ選択
502
文の先頭に特殊記号を付与
503
ポインタを移動
504
ポインタの位置からN文字の文字間接続確率を調べる
505
確率による分割処理
506
ポインタを一文字進める
507
文終了判定
続く
終了
508
次の文を選択
509
文書終了判定
続く
終了
終了

【図６】

(a)

処理対象文字列：　にわにわにわ

↓

文頭特殊記号付与後の
処理対象文字列　：　＃に　に　に

(b)

2グラム出現頻度	
文字列	回数
#に	1
にわ	3
わに	2

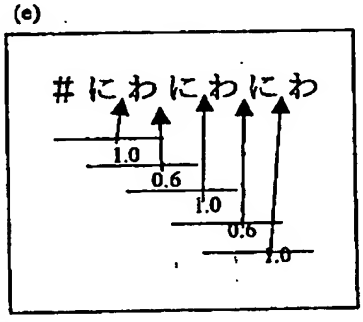
(c)

3グラム出現頻度	
文字列	回数
#にわ	1
にわに	2
わにわ	2

(d)

文字間接続確率		
文字列	確率	計算方法
#にわ	1.0	=Count(#にわ)/Count(#に)
にわに	0.6	=Count(にわに)/Count(にわ)
わにわ	1.0	=Count(わにわ)/Count(わに)

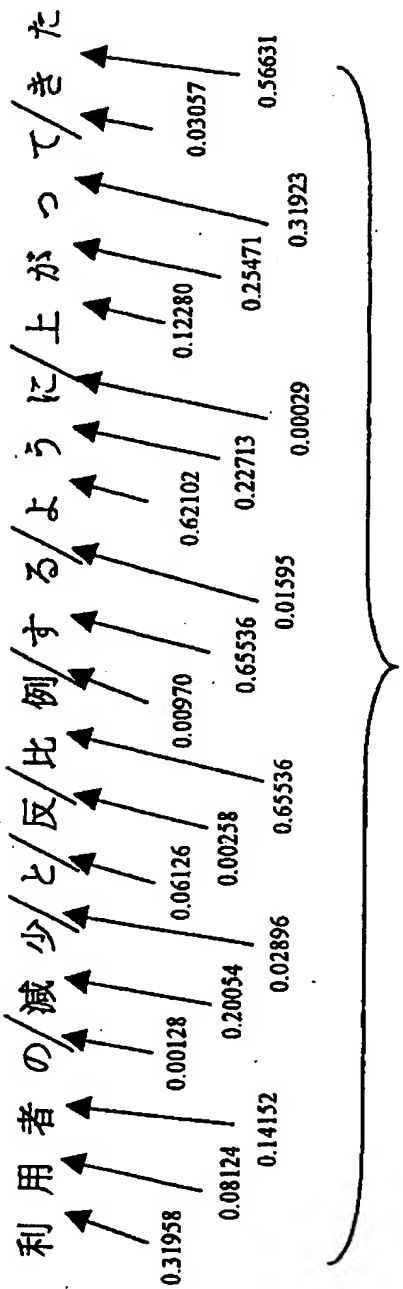
(e)



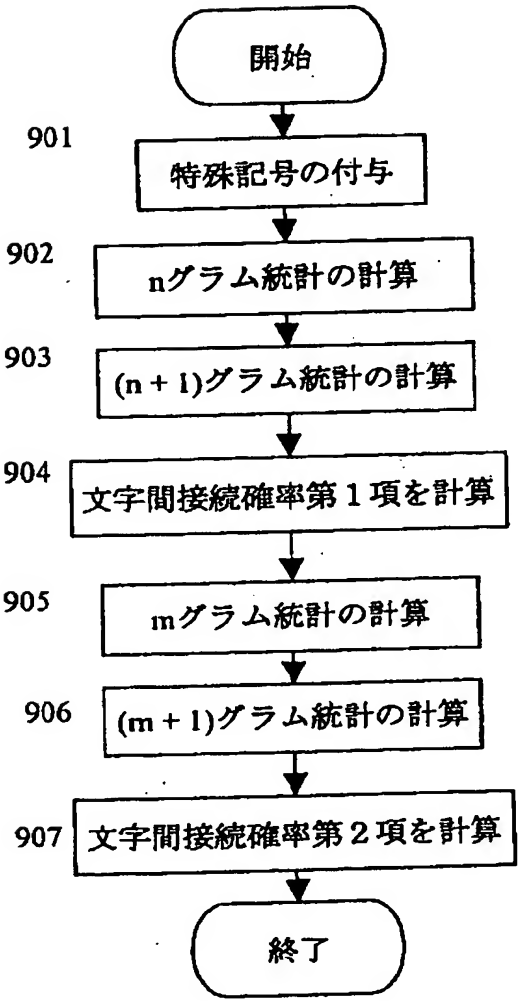
【図７】

文字間接続確率	
文字列	確率
ようち	0.00016
ようっ	0.00011
ようつ	0.00005
ようて	0.00011
ようで	0.01938
ようと	0.08398
ような	0.13665
ように	0.22713
よくあ	0.01593
よくい	0.00612
よくえ	0.00024
よくお	0.00269
よくか	0.00171

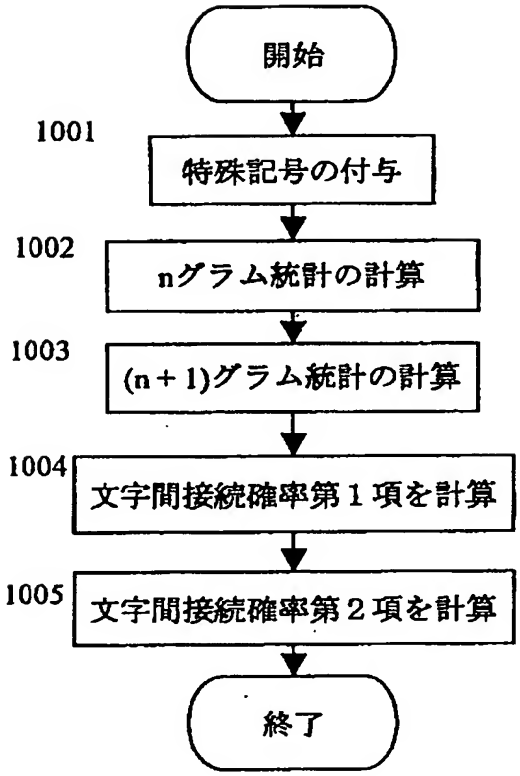
【図８】


利用者の減少と反比例する
0.31958
0.08124
0.14152
0.00128
0.20054
0.02896
0.06126
0.00258
0.65536
0.00970
0.65536
0.01595
0.62102
0.22713
0.00029
0.12280
0.25471
0.31923
0.03057
0.56631

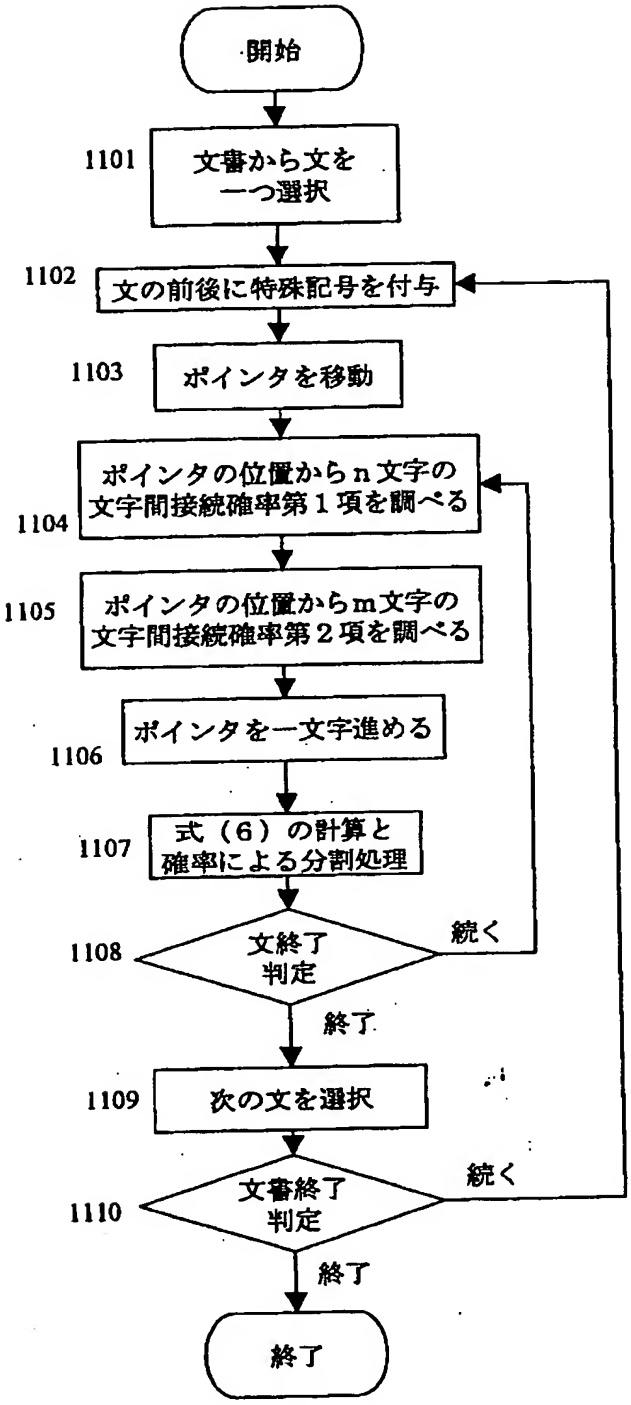
【図９】


開始
901
特殊記号の付与
902
nグラム統計の計算
903
(n + 1)グラム統計の計算
904
文字間接続確率第１項を計算
905
mグラム統計の計算
906
(m + 1)グラム統計の計算
907
文字間接続確率第２項を計算
終了

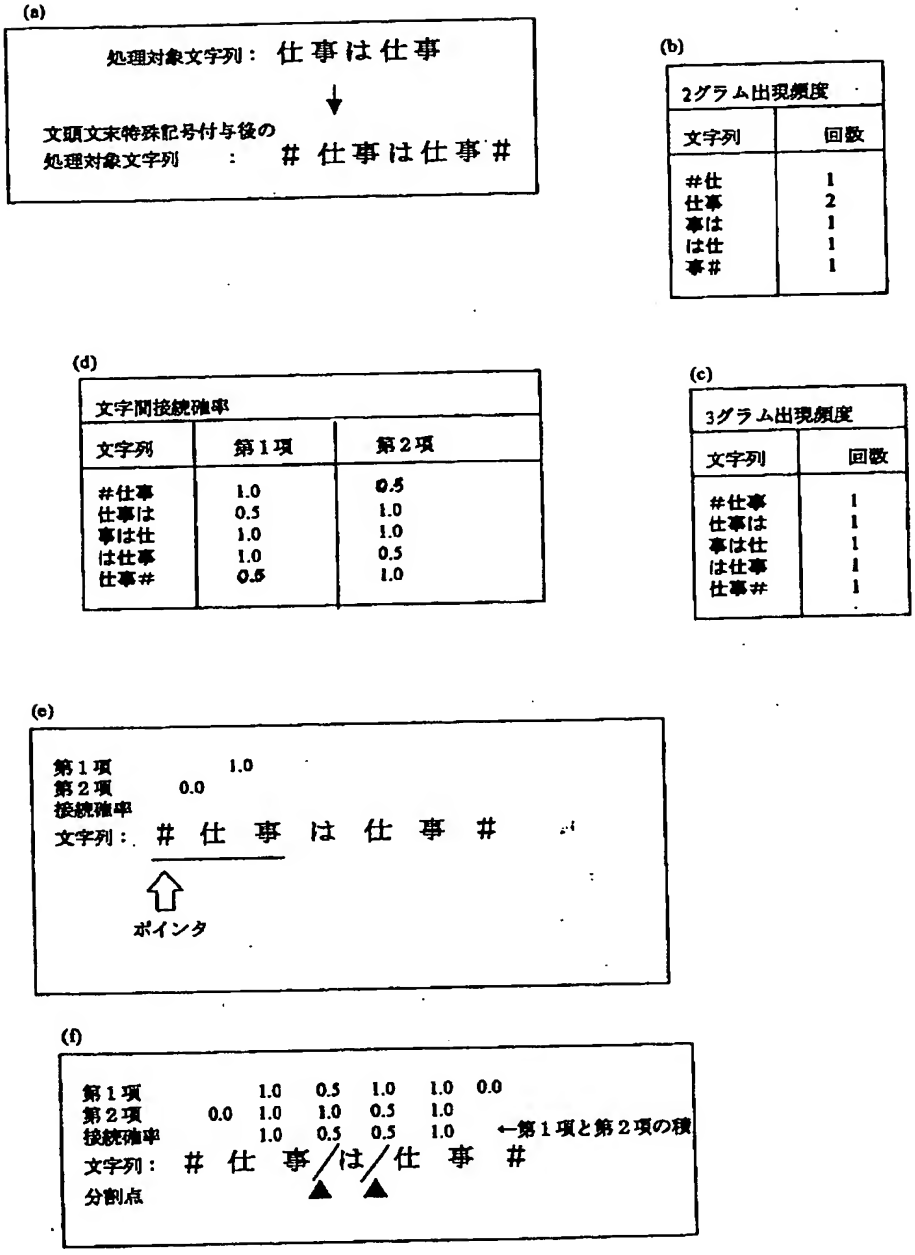
【図１０】


開始
1001
特殊記号の付与
1002
nグラム統計の計算
1003
(n＋1)グラム統計の計算
1004
文字間接続確率第１項を計算
1005
文字間接続確率第２項を計算
終了

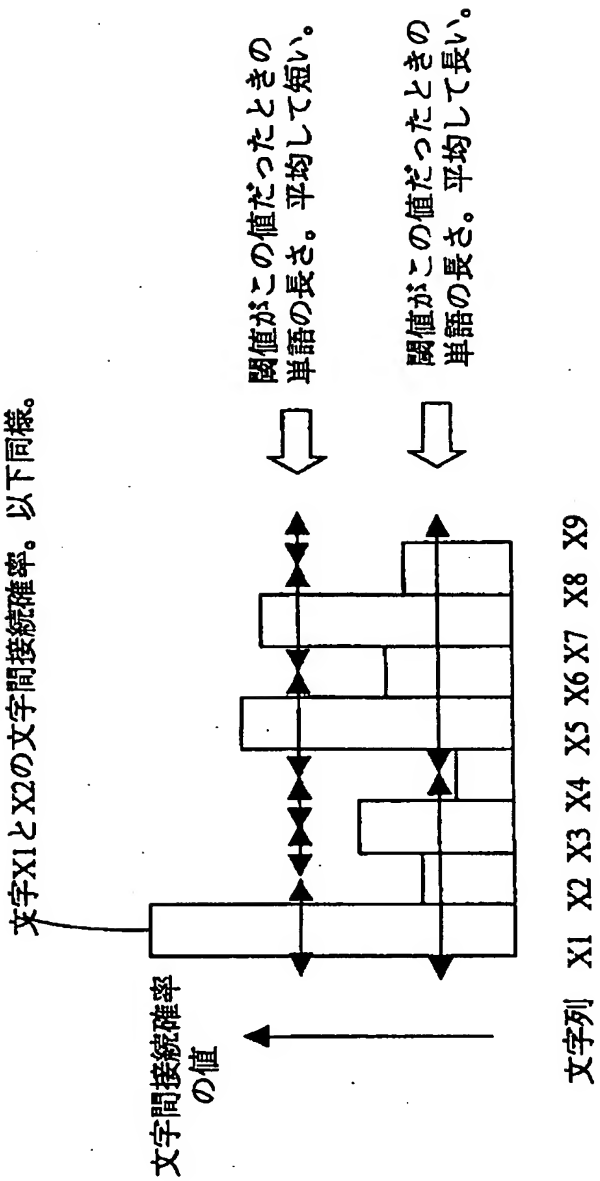
【図１１】


開始
1101
文書から文を一つ選択
1102
文の前後に特殊記号を付与
1103
ポインタを移動
1104
ポインタの位置からｎ文字の文字間接続確率第１項を調べる
1105
ポインタの位置からｍ文字の文字間接続確率第２項を調べる
ポインタを一文字進める
1106
1107
式（６）の計算と確率による分割処理
1108
文終了判定
続く
終了
1109
次の文を選択
1110
文書終了判定
続く
終了
終了

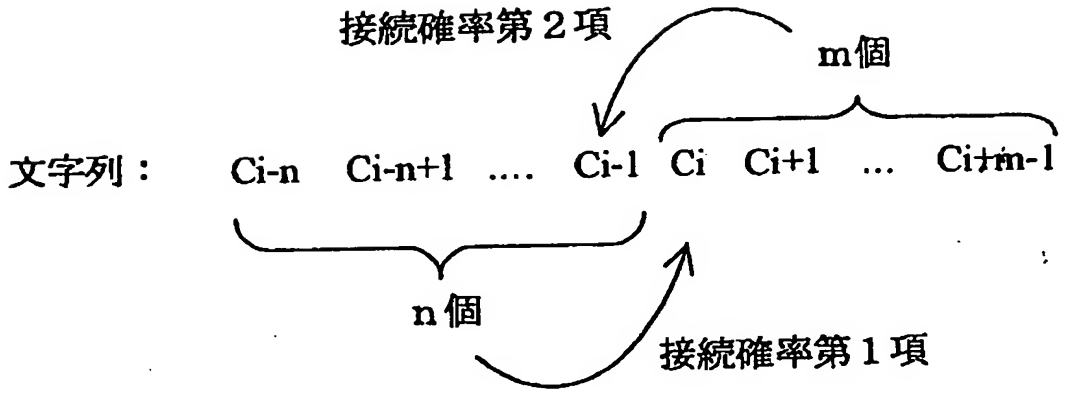
【図１２】


(a)
処理対象文字列：仕事は仕事
文頭文末特殊記号付与後の処理対象文字列：#仕事は仕事#
(b)
2グラム出現頻度
文字列 回数
#仕 1
仕事 2
事は 1
は仕 1
事# 1
(d)
文字間接続確率
文字列 第1項 第2項
#仕事 1.0 0.5
仕事は 0.5 1.0
事は仕 1.0 1.0
は仕事 1.0 0.5
仕事# 0.5 1.0
(c)
3グラム出現頻度
文字列 回数
#仕事 1
仕事は 1
事は仕 1
は仕事 1
仕事# 1
(e)
第1項 1.0
第2項 0.0
接続確率
文字列：# 仕 事 は 仕 事 #
ポインタ
(f)
第1項 1.0 0.5 1.0 1.0 0.0
第2項 0.0 1.0 1.0 0.5 1.0
接続確率 1.0 0.5 0.5 1.0 ←第1項と第2項の積
文字列：# 仕 事 / は / 仕 事 #
分割点

【図１３】

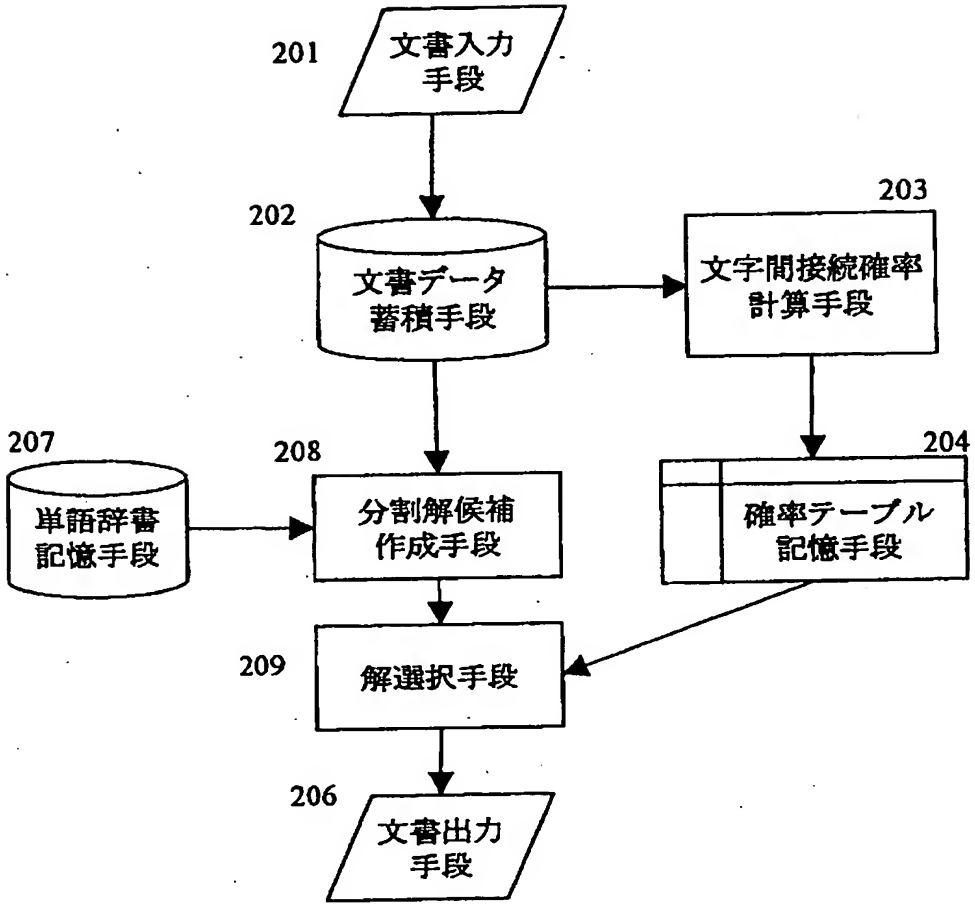

閾値がこの値だったときの
単語の長さ。平均して短い。
閾値がこの値だったときの
単語の長さ。平均して長い。
文字X1とX2の文字間接続確率。以下同様。
文字間接続確率
の値
文字列 X1 X2 X3 X4 X5 X6 X7 X8 X9

【図１４】

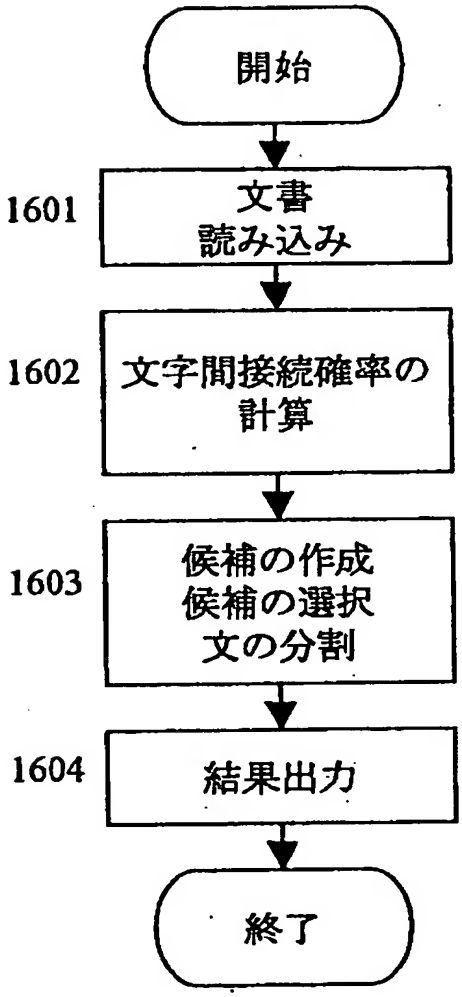


出証特２０００－３１０８２５０

【図１５】

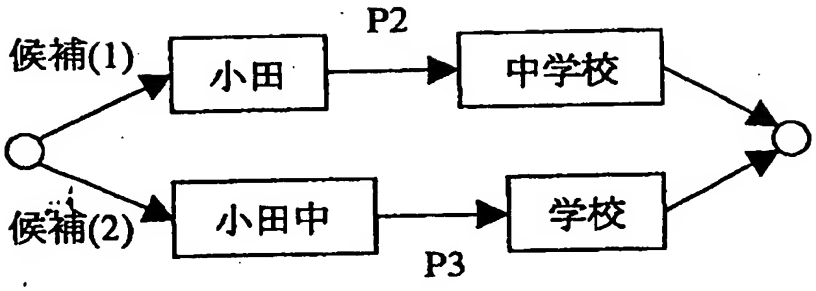

201
文書入力手段
202
文書データ蓄積手段
203
文字間接続確率計算手段
204
確率テーブル記憶手段
207
単語辞書記憶手段
208
分割解候補作成手段
209
解選択手段
206
文書出力手段

【図１６】


開始
1601
文書
読み込み
1602
文字間接続確率の
計算
1603
候補の作成
候補の選択
文の分割
1604
結果出力
終了

【図１７】

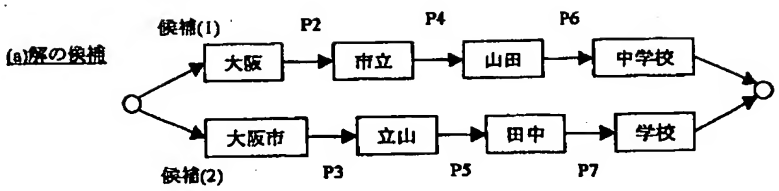
(a)解の候補



(b) 文字間の接続確率表

小	＞	P1 = 0.019
田	＞	P2 = 0.007
中	＞	P3 = 0.088
学	＞	P4 = 0.439
校		

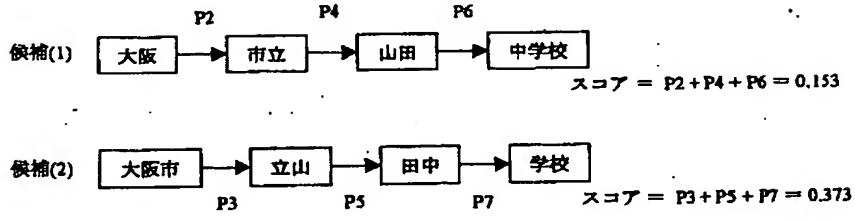
【図１８】



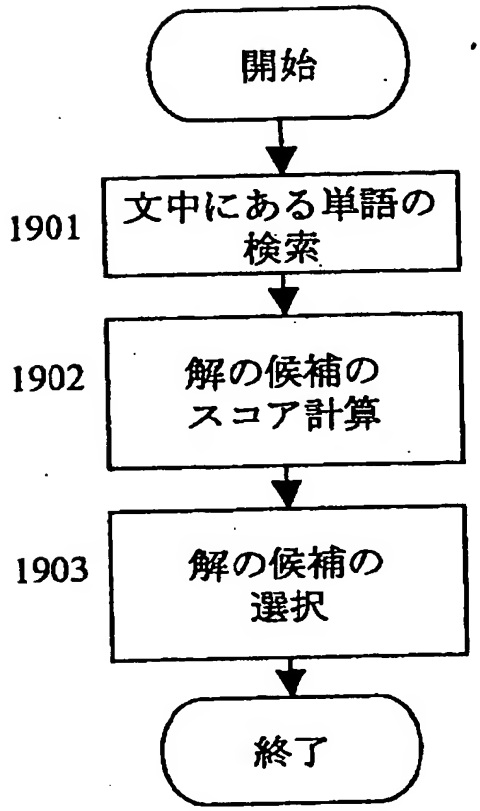
(b) 文字間の接続確率表

大	>	P1＝0.599
阪	>	P2＝0.141
市	>	P3＝0.174
立	>	P4＝0.006
山	>	P5＝0.111
田	>	P6＝0.006
中	>	P7＝0.088
学	>	P8＝0.439
校		

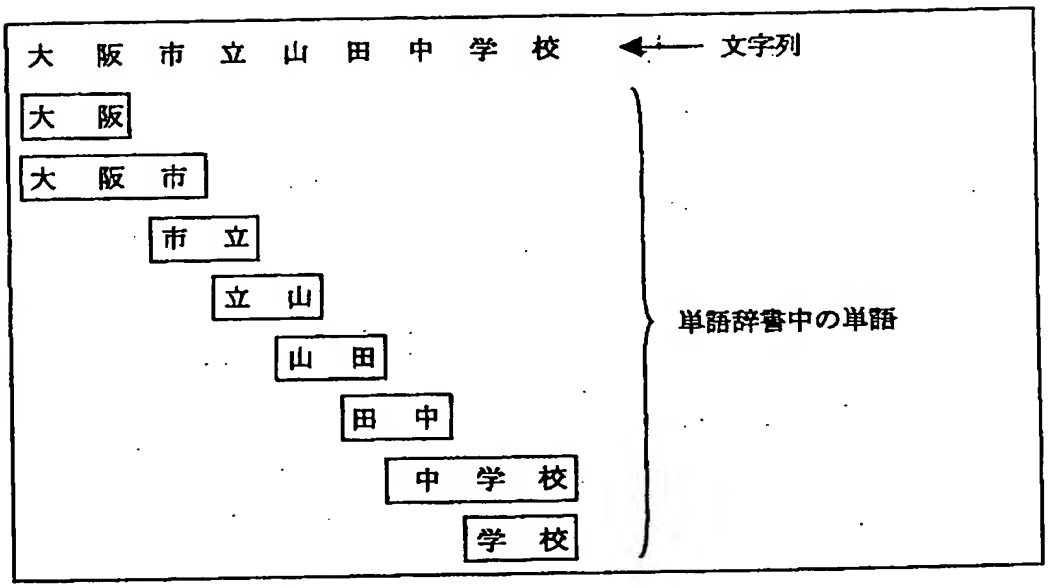
(c)解の候補の比較



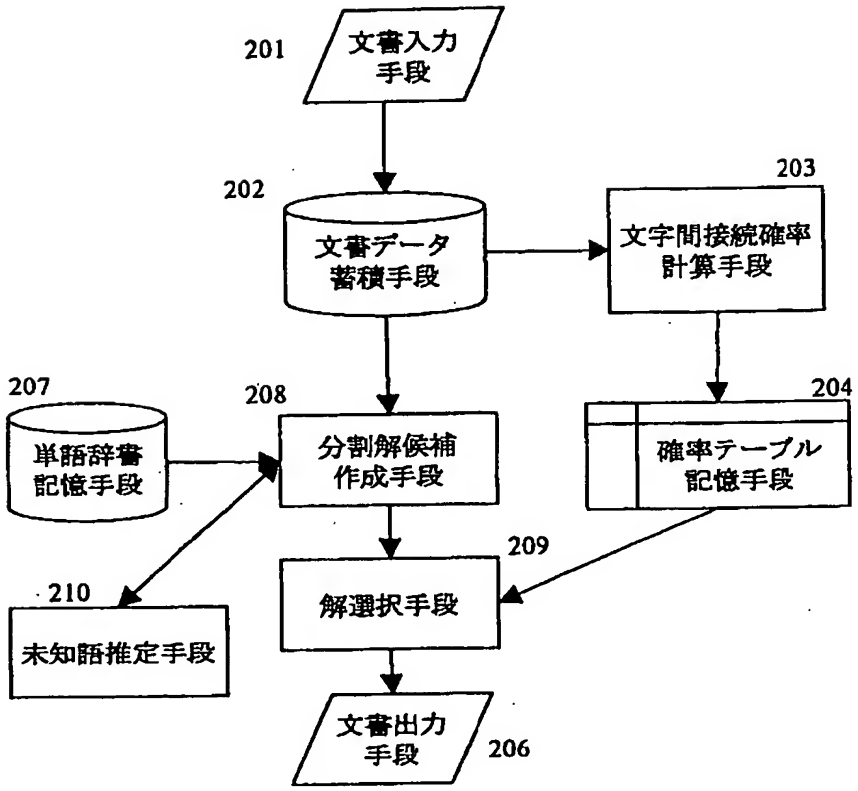
【図１９】



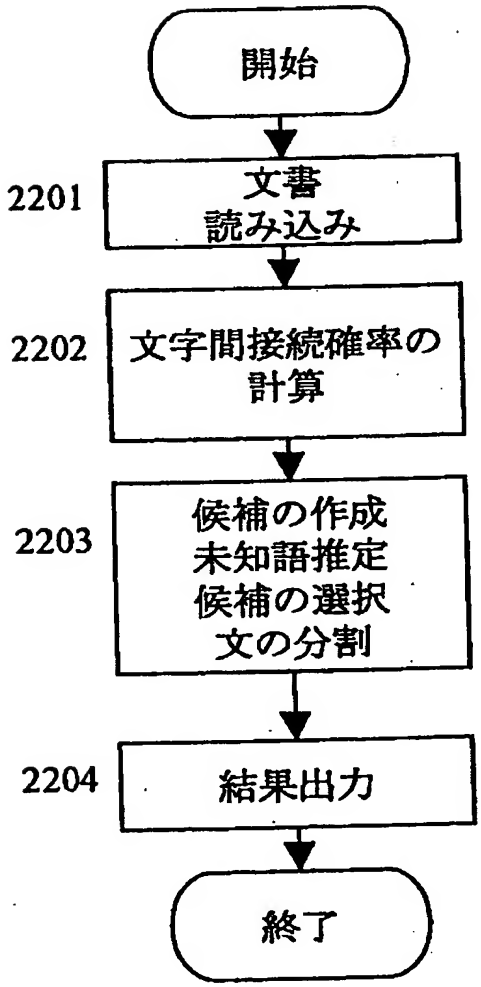
【図２０】



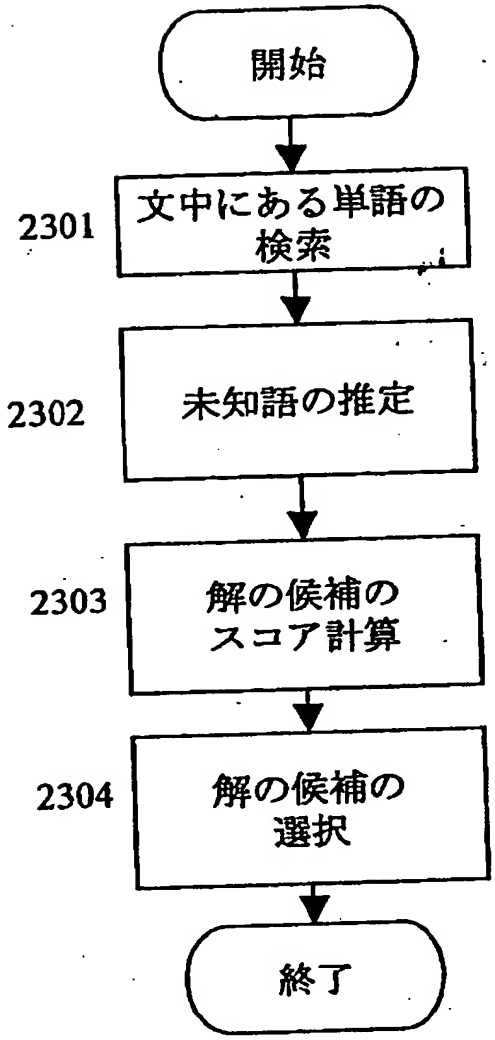
【図２１】


201
文書入力手段
202
文書データ蓄積手段
203
文字間接続確率計算手段
204
確率テーブル記憶手段
207
単語辞書記憶手段
208
分割解候補作成手段
209
解選択手段
210
未知語推定手段
文書出力手段
206

【図２２】

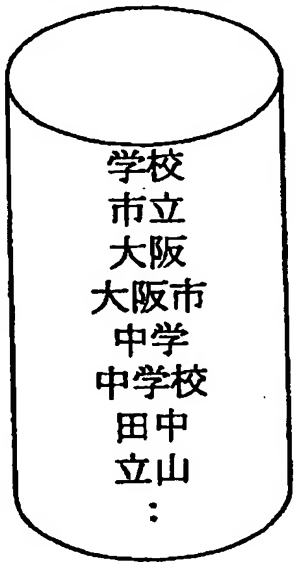

開始
2201
文書
読み込み
2202
文字間接続確率の
計算
2203
候補の作成
未知語推定
候補の選択
文の分割
2204
結果出力
終了

【図２３】

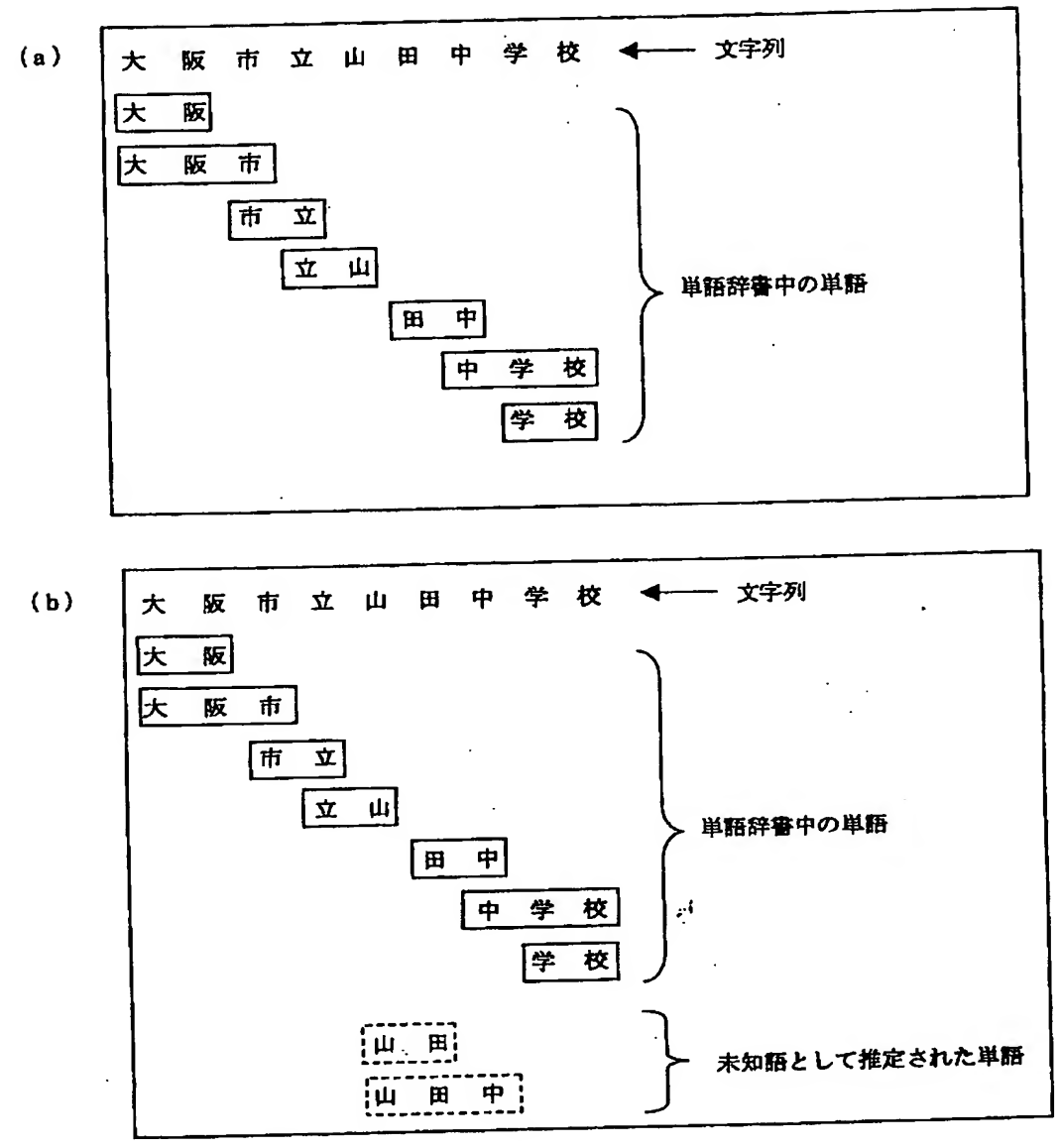

開始
2301
文中にある単語の検索
2302
未知語の推定
2303
解の候補のスコア計算
2304
解の候補の選択
終了

【図２４】

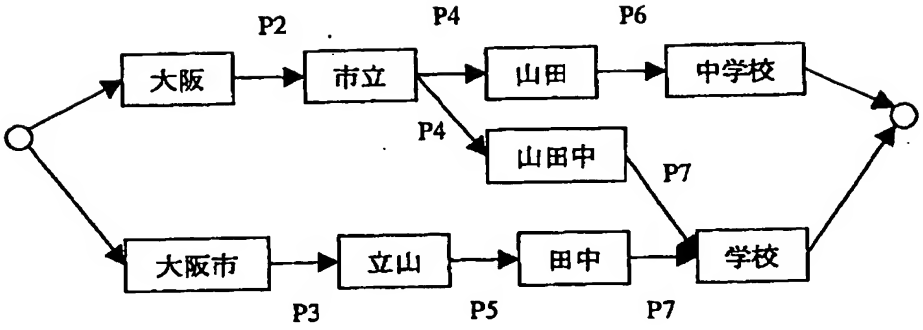
単語辞書記憶手段２０７



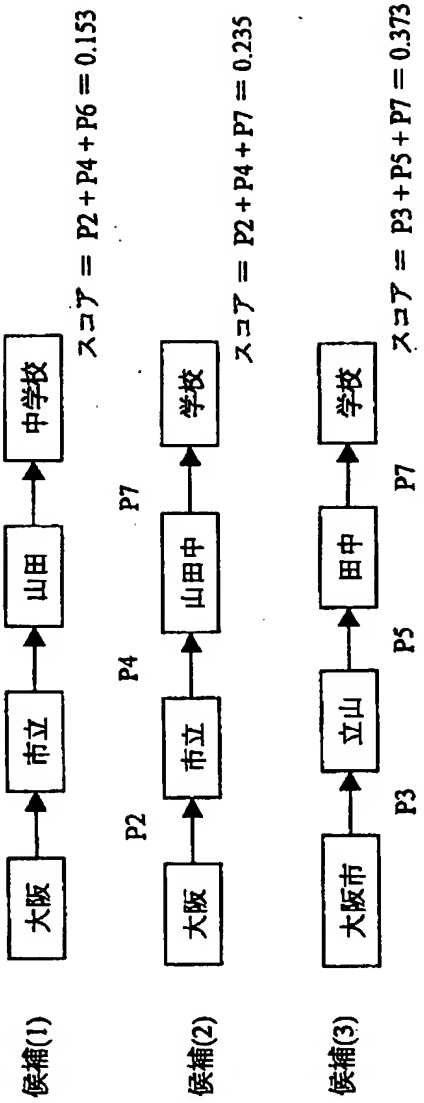
【図２５】


(a)
大 阪 市 立 山 田 中 学 校
文字列
大 阪
大 阪 市
市 立
立 山
田 中
中 学 校
学 校
単語辞書中の単語
(b)
大 阪 市 立 山 田 中 学 校
文字列
大 阪
大 阪 市
市 立
立 山
田 中
中 学 校
学 校
単語辞書中の単語
山 田
山 田 中
未知語として推定された単語

【図２６】

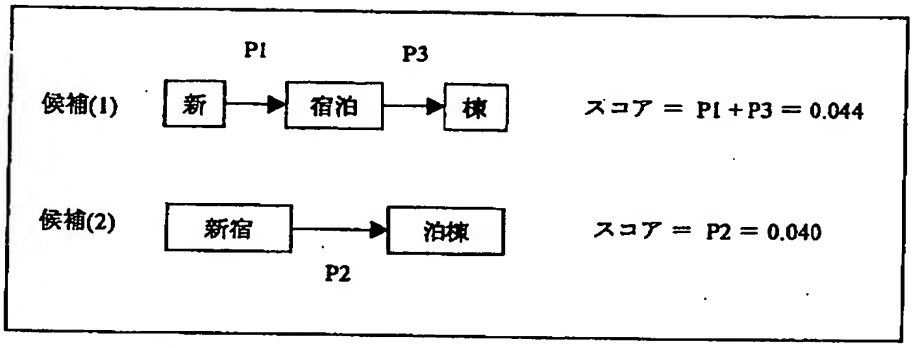


【図２７】

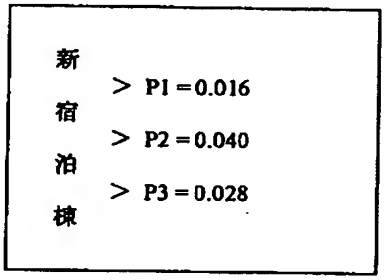

候補(1)
大阪
P2
市立
P4
山田
P6
中学校
スコア = P2 + P4 + P6 = 0.153
候補(2)
大阪
P2
市立
P4
山田中
P7
学校
スコア = P2 + P4 + P7 = 0.235
候補(3)
大阪市
P3
立山
P5
田中
P7
学校
スコア = P3 + P5 + P7 = 0.373

【図２８】

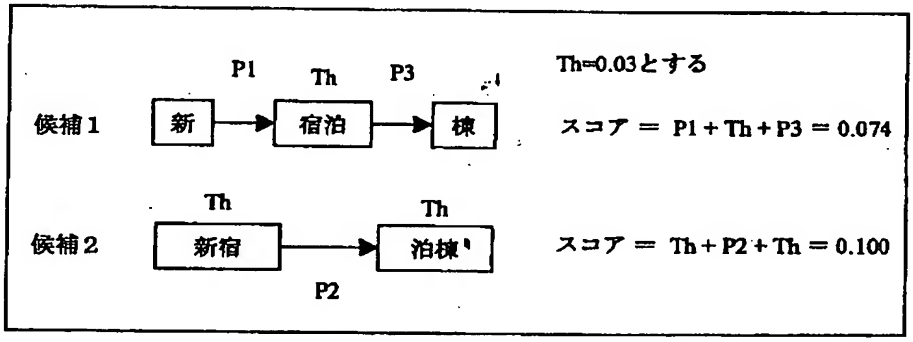
(a)解の候補の比較



(b) 文字間の接続確率表



(c)解の候補の比較



【書類名】　要約書

【要約】

【課題】　電子計算機を利用した自然言語処理システムにおいて、辞書や学習用単語分割済文を必要としない単語分割方式を実現することを目的とする。

【解決手段】　入力された単語分割されていない文書から、文字結合度としての文字間接続確率を統計的に計算し、テーブルに記録する。この文字間接続確率を用いて入力された文書を調べ、文字間接続確率の低い部分で文書を分割し、出力する。

【選択図】　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００５８２１〕

１．変更年月日　　１９９０年　８月２８日
［変更理由］　新規登録
住　所　　大阪府門真市大字門真１００６番地
氏　名　　松下電器産業株式会社